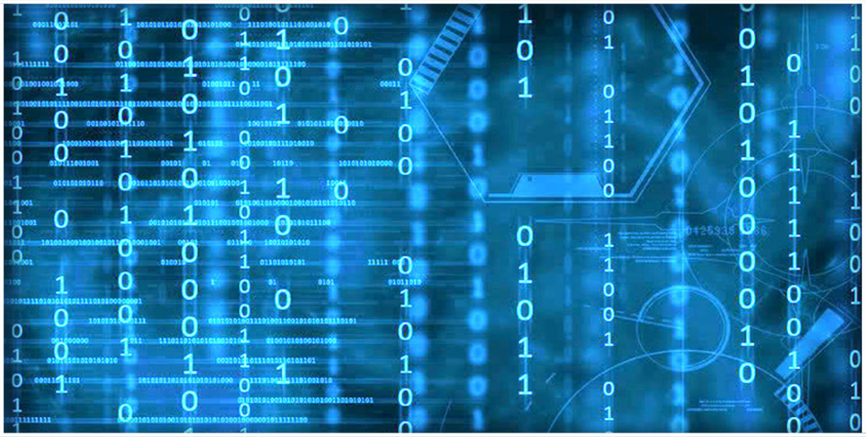
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 10 Issue 8

August 2019



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

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 10 Issue 8 August 2019
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Paper 1: Applying CRISPR-Cas9 Off-Target Editing on DNA based Steganography
Authors: Hong Zhou, Xiaoli Huan
PAGE 1 – 5

Paper 2: A Literature Review on Medicine Recommender Systems
Authors: Benjamin Stark, Constanze Knahl, Mert Aydin, Karim Elish
PAGE 6 – 13

Paper 3: Study and Analysis of Delay Sensitive and Energy Efficient Routing Approach
Authors: Babar Ali, Tariq Mahmood, Muhammad Ayzed Mirza, Saleemullah Memon, Muhammad Rashid, Ekang Francis Ajebesone
PAGE 14 – 20

Paper 4: Analysis of Software Tools for Longitudinal Studies in Psychology
Authors: Pavel Kolyasnikov, Evgeny Nikulchev, Vladimir Belov, Anastasia Silaeva, Alexander Kosenkov, Artem Malykh, Zalina Takhirova, Sergey Malykh
PAGE 21 – 33

Paper 5: An Ontological Model for Generating Complete, Form-based, Business Web Applications
Authors: Daniel Strmečki, Ivan Magdalenić
PAGE 34 – 38

Paper 6: Sound user Interface with Touch Panel for Data and Information Expression and its Application to Meteorological Data Representation
Authors: Kohei Arai
PAGE 39 – 44

Paper 7: Local Average of Nearest Neighbors: Univariate Time Series Imputation
Authors: Anibal Flores, Hugo Tito, Carlos Silva
PAGE 45 – 50

Paper 8: Mortality Prediction based on Imbalanced New Born and Perinatal Period Data
Authors: Wafa M. AlShwaish, Maali Ibr. Alabdulhafith
PAGE 51 – 60

Paper 9: Framework for Digital Data Access Control from Internal Threat in the Public Sector
Authors: Haslidah Halim, Maryati Mohd Yusof
PAGE 61 – 67

Paper 10: Improving Usable-Security of Web based Healthcare Management System through Fuzzy AHP
Authors: Nawaf Rasheed Alharbe
PAGE 68 – 71

Paper 11: Relationship Analysis on the Experience of Hospitalised Paediatric Cancer Patient in Malaysia using Text Analytics Approach
Authors: Zuraini Zainol, Puteri N.E. Nohuddin, Nadhirah Rasid, Hamidah Alias, A. Imran Nordin
PAGE 72 – 79

Paper 12: Prediction of Potential-Diabetic Obese-Patients using Machine Learning Techniques
Authors: Raghda Essam Ali, Hatem El-Kadi, Soha Safwat Labib, Yasmine Ibrahim Saad
PAGE 80 – 88

Paper 13: Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors
Authors: Marwa Almasoud, Tomas E Ward
PAGE 89 – 96

Paper 14: Hadoop MapReduce for Parallel Genetic Algorithm to Solve Traveling Salesman Problem
Authors: Entesar Alanzi, Hachemi Bennaceur
PAGE 97 – 107

Paper 15: Twitter Sentiment Analysis in Under-Resourced Languages using Byte-Level Recurrent Neural Model
Authors: Ridi Ferdiana, Wiliam Fajar, Desi Dwi Purwanti, Artmita Sekar Tri Ayu, Fahim Jatmiko
PAGE 108 – 112

Paper 16: Predicting Return Donor and Analyzing Blood Donation Time Series using Data Mining Techniques
Authors: Anfal Saad Alkahtani, Musfira Jilani
PAGE 113 – 118

Paper 17: Modified Farmland Fertility Optimization Algorithm for Optimal Design of a Grid-connected Hybrid Renewable Energy System with Fuel Cell Storage: Case Study of Ataka, Egypt
Authors: Ahmed A. Zaki Diab, Sultan I. EL-Ajmi, Hamdy M. Sultan, Yahia B. Hassan
PAGE 119 – 132

Paper 18: Smart Rubric-based Systematic Model for Evaluating and Prioritizing Academic Practices to Enhance the Education Outcomes
Authors: Mohammed Al-Shargabi
PAGE 133 – 141

Paper 19: Modeling of the Vegetable Oil Blends Composition
Authors: Olga S. Voskanyan, Igor A. Nikitin, Marina A. Nikitina, Maria V. Klokonos, Daria A. Guseva, Igor V. Zavalishin
PAGE 142 – 147

Paper 20: A Machine Learning Approach towards Detecting Dementia based on its Modifiable Risk Factors
Authors: Reem Bin-Hezam, Tomas E. Ward
PAGE 148 – 158

Paper 21: Most Valuable Player Algorithm for Solving Minimum Vertex Cover Problem
Authors: Hebatullah Khattab, Ahmad Sharieh, Basel A. Mahafzah
PAGE 159 – 167

Paper 22: Non-linear Dimensionality Reduction-based Intrusion Detection using Deep Autoencoder
Authors: S. Sreenivasa Chakravarthi, R. Jagadeesh Kannan
PAGE 168 – 174

Paper 23: Towards a Classification View of Personalized e-Learning with Social Collaboration Support
Authors: Amal Al-Abri, Yassine Jamoussi, Zuhoor AlKhanjari, Naoufel Kraiem
PAGE 175 – 183

Paper 24: An Efficient Design of RPL Objective Function for Routing in Internet of Things using Fuzzy Logic
Authors: Adeeb Saaidah, Omar Almomani, Laila Al-Qaisi, Mohammed Kamel MADI
PAGE 184 – 190

Paper 25: Modelling the Enterprise Architecture Implementation in the Public Sector using HOT-Fit Framework
Authors: Hasimi Sallehudin, Nurhizam Safie Mohd Satar, Nur Azaliah Abu Bakar, Rogis Baker, Farashazillah Yahya, Ahmad Firdause Md Fadzil
PAGE 191 – 198

Paper 26: Robot Arm Analysis based on Master Device Pneumatic Actuators
Authors: Mohd Aliff, Nor Samsiah Sani
PAGE 199 – 204

Paper 27: Attractiveness Analysis of Quiz Games
Authors: Tara Khairiyah Md Zali, Nor Samsiah Sani, Abdul Hadi Abd Rahman, Mohd Aliff
PAGE 205 – 210

Paper 28: A Survey: Agent-based Software Technology Under the Eyes of Cyber Security, Security Controls, Attacks and Challenges
Authors: Bandar Alluhaybi, Mohamad Shady Alrahhal, Ahmed Alzhrani, Vijey Thayananthan
PAGE 211 – 230

Paper 29: Indoor Positioning System using Regression-based Fingerprint Method
Authors: Reginald Putra Ghozali, Gede Putra Kusuma
PAGE 231 – 239

Paper 30: A New Security Model for Web Browser Local Storage
Authors: Thamer Al-Rousan, Bassam Al-Shargabi, Hasan Abualese
PAGE 240 – 247

Paper 31: A Novel Approach for Dimensionality Reduction and Classification of Hyperspectral Images based on Normalized Synergy
Authors: Asma Elmaizi, Hasna Nhaila, Elkebir Sarhrouni, Ahmed Hammouch, Nacir Chafik
PAGE 248 – 257

Paper 32: Muscle Electro Stimulator for the Reduction of Stretch Marks
Authors: Paima-Sahuma Jaime, Duran-Berrio Linder, Palacios-Cusiyunca Chrisstopher, Roman-Gonzalez Avid
PAGE 258 – 263

Paper 33: Robust Video Content Authentication using Video Binary Pattern and Extreme Learning Machine
Authors: Mubbashar Sadddique, Khurshid Asghar, Tariq Mehmood, Muhammad Hussain, Zulfiqar Habib
PAGE 264 – 269

Paper 34: Automated Greenhouses for the Reduction of the Cost of the Family Basket in the District of Villa El Salvador-Perú
Authors: Pedro Romero Huaroto, Abraham Casanova Robles, Nicolh Antony Ciriaco Susanibar, Avid Roman-Gonzalez
PAGE 270 – 274

Paper 35: Learning Analytics Framework for Adaptive E-learning System to Monitor the Learner's Activities
Authors: Salma EL Janati, Abdelilah Maach, Driss El Ghanami
PAGE 275 – 284

Paper 36: Machine Learning Approaches for Predicting the Severity Level of Software Bug Reports in Closed Source Projects

Authors: Aladdin Baarah, Ahmad Aloqaily, Zaher Salah, Mannam Zamzeer, Mohammad Sallam

PAGE 285 – 294

Paper 37: Steganography Performance over AWGN Channel

Authors: Fahd Alharbi

PAGE 295 – 299

Paper 38: Cloud Security based on the Homomorphic Encryption

Authors: Waleed T. Al-Sit, Qussay Al-Jubouri, Hani Al-Zoubi

PAGE 300 – 307

Paper 39: Video Analysis with Faces using Harris Detector and Correlation

Authors: Rodolfo Romero Herrera, Francisco Gallegos Funes, José Elias Romero Martínez

PAGE 308 – 313

Paper 40: Mapping of Independent Tasks in the Cloud Computing Environment

Authors: Biswajit Nayak, Sanjay Kumar Padhi

PAGE 314 – 318

Paper 41: Convolutional Neural Network Architecture for Plant Seedling Classification

Authors: Heba A. Elnemr

PAGE 319 – 325

Paper 42: A Methodology for Engineering Domain Ontology using Entity Relationship Model

Authors: Muhammad Ahsan Raza, M. Rahmah, Sehrish Raza, A. Noraziah, Roslina Abd. Hamid

PAGE 326 – 332

Paper 43: An Evaluation Model for Auto-generated Cognitive Scripts

Authors: Ahmed M. ELMougi, Rania Hodhod, Yasser M. K. Omar

PAGE 333 – 340

Paper 44: Systematic Literature Review of Identifying Issues in Software Cost Estimation Techniques

Authors: Muhammad Asif Saleem, Rehan Ahmad, Tahir Alyas, Muhammad Idrees, Asfandayar, Asif Farooq, Adnan Shahid Khan, Kahawaja Ali

PAGE 341 – 346

Paper 45: A Novel Student Clustering Model for the Learning Simplification in Educational Environments

Authors: Khalaf Khatatneh, Islam Khataleen, Rami Alshwaiyat, Mohammad Wedyan

PAGE 347 – 351

Paper 46: Classifying Cardiotocography Data based on Rough Neural Network

Authors: Belal Amin, Mona Gamal, A. A. Salama, I.M. El-Henawy, Khaled Mahfouz

PAGE 352 – 356

Paper 47: Hospital Queue Control System using Quick Response Code (QR Code) as Verification of Patient's Arrival

Authors: Ridho Hendra Yoga Perdana, Hudiono, Mochammad Taufik, Amalia Eka Rakhmania, Rohman Muhamad Akbar, Zainul Arifin

PAGE 357 – 363

Paper 48: Artificial Immune-based Algorithm for Academic Leadership Assessment

Authors: Hamidah Jantan, Nur Hamizah Syafiqah Che Azemi, Zulaiha Ali Othman

PAGE 364 – 370

Paper 49: Antennas of Circular Waveguides

Authors: Cusacani Guerrero, Julio Agapito, Roman-Gonzalez Avid

PAGE 371 – 374

Paper 50: Pedestrian Crossing Safety System at Traffic Lights based on Decision Tree Algorithm

Authors: Denny Hardiyanto, Iswanto, Dyah Anggun Sartika, Muamar Rojali

PAGE 375 – 379

Paper 51: Autonomous Monitoring System using Wi-Fi Economic

Authors: Michael Ames Ccoa Garay, Avid Roman-Gonzalez

PAGE 380 – 386

Paper 52: Impact of ICT on Students' Academic Performance: Applying Association Rule Mining and Structured Equation Modeling

Authors: Mohammad Aman Ullah, Mohammad Manjur Alam, Ahmed Shan-A-Alahi, Mohammed Mahmudur Rahman, Abdul Kadar Muhammad Masum, Nasrin Akter

PAGE 387 – 393

Paper 53: WhatsApp as an Educational Support Tool in a Saudi University

Authors: Ahmad J Reeves, Salem Alkhalaf, Mohamed A. Amasha

PAGE 394 – 401

Paper 54: Real-Time Intelligent Parking Entrance Management

Authors: Sofia Belkhala, Siham Benhadou, Hicham Medromi

PAGE 402 – 405

Paper 55: LUCIDAH Ligative and Unligative Characters in a Dataset for Arabic Handwriting

Authors: Yousef Elarian, Irfan Ahmad, Abdelmalek Zidouri, Wasfi G. Al-Khatib

PAGE 406 – 415

Paper 56: An Efficient Normalized Restricted Boltzmann Machine for Solving Multiclass Classification Problems

Authors: Muhammad Aamir, Nazri Mohd Nawi, Fazli Wahid, Hairulnizam Mahdin

PAGE 416 – 426

Paper 57: Key Schedule Algorithm using 3-Dimensional Hybrid Cubes for Block Cipher

Authors: Muhammad Faheem Mushtaq, Sapiee Jamel, Siti Radhiah B. Megat, Urooj Akram, Mustafa Mat Deris

PAGE 427 – 442

Paper 58: Web Service Testing Techniques: A Systematic Literature Review

Authors: Israr Ghani, Wan M.N. Wan-Kadir, Ahmad Mustafa

PAGE 443 – 458

Paper 59: Deep Transfer Learning Application for Automated Ischemic Classification in Posterior Fossa CT Images

Authors: Anis Azwani Muhd Suberi, Wan Nurshazwani Wan Zakaria, Razali Tomari, Ain Nazari, Mohd Norzali Hj Mohd, Nik Farhan Nik Fuad

PAGE 459 – 465

Paper 60: Effect of e-Commerce Platforms towards Increasing Merchant's Income in Malaysia
Authors: M.Hafiz Yusoff, Mohammad Ahmed Alomari, Nurul Adilah Abdul Latiff, Motea S. Alomari
PAGE 466 – 470

Paper 61: Novel Adaptive Auto-Correction Technique for Enhanced Fingerprint Recognition
Authors: Thejaswini P, Srikantaswamy R S, Manjunatha A S
PAGE 471 – 479

Paper 62: Empirical Study of Segment Particle Swarm Optimization and Particle Swarm Optimization Algorithms
Authors: Mohammed Adam Kunna Azrag, Tuty Asmawaty Abdul Kadir
PAGE 480 – 485

Paper 63: An Efficient Deep Learning Model for Olive Diseases Detection
Authors: Madallah Alruwaili, Saad Alanazi, Sameh Abd El-Ghany, Abdulaziz Shehab
PAGE 486 – 492

Paper 64: Biotechnical System for Recording Phonocardiography
Authors: Marwan Ahmed Ahmed Hamid, Maria Abdullah, Najeed Ahmed Khan, Yasmin Mohammed Ahmed AL-Zoom
PAGE 493 – 497

Paper 65: A Guideline for Decision-making on Business Intelligence and Customer Relationship Management among Clinics
Authors: Nur Izzati Yusof, Norziha Megat Mohd. Zainuddin, Noor Hafizah Hassan, Nilam Nur Amir Sjarif, Suraya Yaacob, Wan Azlan Wan Hassan
PAGE 498 – 505

Paper 66: Cognitive Neural Network Classifier for Fault Management in Cloud Data Center
Authors: S.Indirani, C.Jothi Venkateswaran
PAGE 506 – 511

Paper 67: Deep Learning Classification of Biomedical Text using Convolutional Neural Network
Authors: Rozilawati Dollah, Chew Yi Sheng, Norhawaniah Zakaria, Mohd Shahizan Othman, Abd Wahid Rasib
PAGE 512 – 517

Paper 68: Acquisition and Classification System of EMG Signals for Interpreting the Alphabet of the Sign Language
Authors: Alvarado-Diaz Witman, Meneses-Claudio Brian, Fiorella Flores-Medina, Patricia Condori, Natalia I. Vargas-Cuentas, Avid Roman-Gonzalez
PAGE 518 – 521

Paper 69: Implementation of a Beowulf Cluster and Analysis of its Performance in Applications with Parallel Programming
Authors: Enrique Lee Huamaní, Patricia Condori, Avid Roman-Gonzalez
PAGE 522 – 527

Paper 70: An Internet of Things (IOT) based Smart Parking Routing System for Smart Cities
Authors: Elgarej Mouhcine, Karouani Yassine, El Fazazi Hanaa, Khalifa Mansouri, Youssfi Mohamed
PAGE 528 – 538

Paper 71: An Automated Approach for Identification of Non-Functional Requirements using Word2Vec Model
Authors: Muhammad Younas, Karzan Wakil, Dayang N. A. Jawawi, Muhammad Arif Shah, Ahmad Mustafa
PAGE 539 – 547

Paper 72: VoIP QoS Analysis over Asterisk and Axon Servers in LAN Environment

Authors: Naveed Ali Khan, Abdul Sattar Chan, Kashif Saleem, Zuhaibuddin Bhutto, Ayaz Hussain

PAGE 548 – 554

Paper 73: Privacy Preserving Data Mining Approach for IoT based WSN in Smart City

Authors: Ahmed M. Khedr, Walid Osamy, Ahmed Salim, Abdel-Aziz Salem

PAGE 555 – 563

Paper 74: Efficient Distributed SPARQL Queries on Apache Spark

Authors: Saleh Albahli

PAGE 564 – 568

Paper 75: A Modular Aspect-Oriented Programming Approach of Join Point Interfaces

Authors: Cristian Vidal, Erika Madariaga, Claudia Jim´enez, Luis Carter

PAGE 569 – 574

Paper 76: Artificial Potential Field Algorithm Implementation for Quadrotor Path Planning

Authors: Iswanto, Alfian Ma'arif, Oyas Wahyunggoro, Adha Imam Cahyadi

PAGE 575 – 585

Paper 77: New Transport Layer Security using Metaheuristics and New Key Exchange Protocol

Authors: Mohamed Kaddouri, Mohammed Bouhdadi, Zakaria Kaddouri, Driss Guerchi, Bouchra Echandouri

PAGE 586 – 591

Paper 78: Arabic Lexicon Learning to Analyze Sentiment in Microblogs

Authors: Mahmoud B. Rokaya, Ahmed S. Ghiduk, Mahmoud B. Rokaya, Ahmed S. Ghiduk

PAGE 592 – 599

Paper 79: An Efficient Segmentation of Retinal Blood Vessels using Singular Value Decomposition and Morphological Operator

Authors: N. C. Santosh Kumar, Y. Radhika

PAGE 600 – 609

Paper 80: A Survey of Various Frameworks and Solutions in all Branches of Digital Forensics with a Focus on Cloud Forensics

Authors: Mohammed Khanafseh, Mohammad Qatawneh, Wesam Almobaideen

PAGE 610 – 629

Applying CRISPR-Cas9 Off-Target Editing on DNA based Steganography

Hong Zhou[1]
Department of Mathematical Sciences
University of Saint Joseph
West Hartford, Connecticut, USA

Xiaoli Huan[2]
Department of Computer Science
Troy University
Troy, Alabama, USA

***Abstract*—Different from cryptography which encodes data into an incomprehensible format difficult to decrypt, steganography hides the trace of data and therefore minimizes attention to the hidden data. To hide data, a carrier body must be utilized. In addition to the traditional data carriers including images, audios, and videos, DNA emerges as another promising data carrier due to its high capacity and complexity. Currently, DNA based steganography can be practiced with either biological DNA substances or digital DNA sequences. In this article, we present a digital DNA steganography approach that utilizes the CRISPR-Cas9 off-target editing such that the secret message is fragmented into multiple sgRNA homologous sequences in the genome. Retrieval of the hidden message mimics the Cas9 off-target detection process which can be accelerated by computer software. The feasibility of this approach is analyzed, and practical concerns are discussed.**

***Keywords*—DNA; steganography; CRISPR; Cas9; sgRNA; off-target; substitution**

I. Introduction

Steganography is a technology that integrates the secret data into a common message seamlessly to avoid any attention to the data (note that data and message are two exchangeable concepts in this article). It is different from cryptography. If the data need to be secured from decryption, then cryptography technology is required. The power of cryptography is that even the encrypted message is left on the open ground, the meaning of the message cannot be revealed without the proper key. Its drawback lies in the fact that the data can catch attention. In contrast, steganography hides the message in a data carrier and makes minimal modifications to the carrier so that most people won't imagine any secret messages inside. However, hiding plain text messages via steganography is not recommended given the great advances in digital technologies. Today, steganography is used to place another layer of protection above cryptography in many applications.

Traditionally steganography makes use of either image, audio or video media. One of the most adopted steganography techniques is the Least Significant Bit (LSB) image steganography in which the least significant bits of the pixels of the cover image embed the secret message. The difference between the cover image and the stego-image (the modified image carrying the secret message) is imperceptible to human visual system [1]. Due to its complexity and large capacity, DNA, either digital DNA sequence, or DNA substance, is becoming another promising data carrier [2, 3, 4]. For example, human genome is about three billion base pairs which can store a large amount of information, making it a powerful data carrier in DNA based steganography.

II. Related Work

A single CPU of modern electronic computer works in a linear fashion, which limits its power in solving NP-complete problems when the size of the problem becomes large, for instance, the directed Hamiltonian path problem. However, in 1994, Leonard Adleman demonstrated that by applying step-by-step DNA biochemical reactions, a process like a computer science algorithm, the directed Hamiltonian path problem could be solved efficiently due to the large number of simultaneous DNA reactions [5]. Since then, DNA computing has become an eye-catching branch in computer science [2, 3, 6, 7, 8, 9, 10, 11, 12]. In 1999, Clelland et al. encoded a secret message as a DNA fragment (the secret-message DNA) and hid it among many other "junk" DNA fragments. Such DNA samples can be prepared as microdots to be delivered to recipients who can only retrieve the secret message by applying polymerase chain reaction (PCR) on the DNA sample with the knowledge of the two PCR primers followed by DNA sequencing [2]. This work started DNA based steganography. However, one drawback of this technique is that the recipient must be given the primer information. Leaking of the primer information can cause data insecurity. A recent study adds one extra layer of protection to the PCR primer by applying the trans-cleavage activity of CRISPR-Cas12a nuclease [13]. Note that CRISPR stands for Clustered Regularly Interspaced Short Palindromic Repeats, and Cas stands for CRISPR Associated System. In this study, fake primers or redundant DNA sequences were pre-ligated to the real PCR primers so that the real primers are concealed. To obtain the real primers, the recipient must apply CRISPR-Cas12a to cut off the fake primers or the redundant sequences [13].

A critical disadvantage of hiding the secret message into a biological DNA sample is the difficulty in preparing the sample and retrieving the secret message through a series of biological experiments. Such experiments are likely to be proceeded by experienced workers in labs armed with modern and expensive equipment. In addition, carrying biological samples may be considered illegal under some circumstances. Steganography based on digital DNA sequences however, provides much better feasibility.

There are three classical methods to hide data into DNA sequences, namely the insertion method, the complementary pair method, and the substitution method [4]. Several revisions of these methods were proposed, but the fundamental techniques keep the same [7, 9, 11]. All the methods require a binary coding rule to convert between binary digits and nucleotide bases (A, C, G, T). A key requirement of the substitution method is that a nucleotide can be substituted by another specific nucleotide, and the mapping is one-to-one [4]. Another requirement is that the length of the secret message must be no longer than the reference DNA (the data carrier) [4].

In the substitution method, the reference sequence R is known to the sender and the recipient along with many other public reference sequences. In the basic scenario, the message M is converted into a binary string and each bit is randomly but sequentially stored by a user-defined substitution function. The modified DNA sequence R’ is delivered to the recipient. The recipient retrieves the secret message by comparing R and R’ through the substitution function. The cracking of the substitution method requires the knowledge of the reference sequence R and the substitution function, and the cracking probability by a random guess is computed as $\frac{1}{6N}$, where N is the number of public reference sequences available [4]. Note that there are six different one-to-one substitution functions available [4], and currently there are about one billion publicly available reference DNA sequences [14].

This article presents an improved substitution method that can render a much lower cracking probability. This method is inspired by the ground-breaking biotechnology CRISPR-Cas9. It simulates the bacteria adaptive immune system and the evolutional arm-races between bacteria and phages.

CRISPR-Cas9 has become a widely adopted tool capable of gene editing, gene expression suppression/activation, and epigenetic modifications [15, 16]. Its gene editing function can usually be achieved by two approaches respectively. One approach introduces frame-shift insertions and/or deletions inside a targeted gene to generate gene-specific knock-outs. The success of this application depends on the error-prone non-homologous end joining (NHEJ) pathway that repairs the double strand DNA break introduced by the Cas9 nuclease. During the NHEJ repairing process, various mutations can be randomly generated and a few of them may result in loss-of-function knockouts [17]. The second approach relies on base-editor platforms which do not generate any double strand DNA breaks [18]. The beauty of CRISPR-Cas9 lies in the fact that it is programmable [15]. The core of CRISPR-Cas9 system includes the Cas9 endonuclease and a single guide RNA (sgRNA). sgRNA has two sequence regions, one is the scaffold to which Cas9 binds, the other is the spacer sequence of about 20 bases. The spacer sequence can be user defined and it can lead the Cas9 to any a genomic locus where the target DNA sequence matches the spacer and has a protospacer adjacent motif (PAM) immediately downstream [15, 19]. The PAM sequences vary depending on the types of Cas nucleases. The most commonly used Cas protein is SpCas9 whose PAM sequences are NGG (AGG, CGG, GGG, and TGG). A critical concern in the CRISPR-Cas9 technology is however, the sgRNA can also lead the Cas9 to other genomic loci that share sequence similarity with the sgRNA spacer, causing off-target genome editing [19]. This scenario is explained in Fig. 1 and 2.

It is understood that Cas9 off-target nuclease activity is likely a result of the evolutional arm-races between bacteria and infection virus. Bacteria that survived virus infection store characteristic virus genetic sequences as spacers between repeated sequences. These spacers can be transcribed into RNA to guide the Cas nuclease to degrade virus genetic elements containing the same sequence. This process provides bacteria an adaptive immune system to resist repeated phage invasion. In response to the selection pressure of the host bacteria, the phages’ genetic sequence evolves with variation(s) to escape Cas nuclease degradation. In return, bacteria CRISPR-Cas system evolves by allowing Cas nuclease to degrade sequences sharing certain degrees of homology with the spacer sequence.

The larger a genome’s size, the more the potential off-target loci for a given sgRNA spacer sequence. For genomes as large as the human genome, identifying the off-target sites becomes a time-consuming process. Different computational tools have been developed to help predict potential genome off-target loci [20, 21]. The early computational methods are mostly built upon the found sgRNA sequence features regarding SpCas9 [15, 22, 23]. However, some discoveries from different research groups are not in agreement with each other, though certain general understandings have reached consensus. 1) Off-target effect decreases when the number of mismatches (including both base mismatches and bulges) between sgRNA and target sequence increases; 2) Cas9 is less tolerant with mismatches proximal to PAM. Later methods incorporate Cas9 domain knowledge, especially energetics parameters, and therefore can achieve better predication results [24].

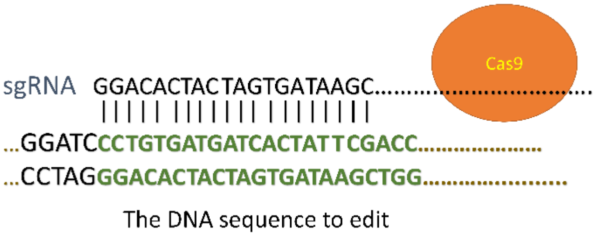

Cas9
sgRNA GGACACTACTAGTGATAAGC……………………………
| | | | | | | | | | | | | | | | | | | |
…GGATCCCTGTGATGATCACTATTCGACC…………………
…CCTAGGGACACTACTAGTGATAAGCTGG……………………
The DNA sequence to edit

Fig. 1. CRISPR-Cas9 on-Target Editing in which the sgRNA Sequence has a Perfect Match with the DNA Sequence.

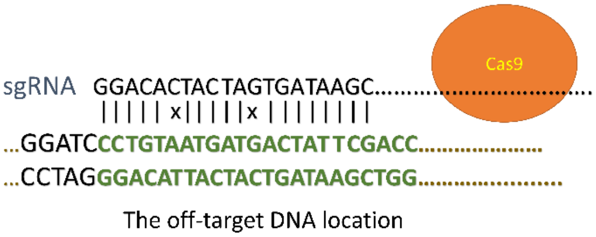

Cas9
sgRNA GGACACTACTAGTGATAAGC……………………………
| | | | | x | | | | | x | | | | | | | |
…GGATCCCTGTAATGATGACTATTCGACC…………………
…CCTAGGGACATTACTACTGATAAGCTGG……………………
The off-target DNA location

Fig. 2. CRISPR-Cas9 off-Target Editing in which the sgRNA Sequence is Homologous to the DNA Sequence.

III. Proposed Algorithm

Our proposed algorithm is based on the following assumptions regarding off-target homology search:

1) All off-target sites must have a primary NGG PAM immediately downstream the sgRNA spacer binding location.

2) All off-target sites can have up to five base mismatches within a given sgRNA spacer sequence. If there are at least six base mismatches, the DNA sequence in study is not considered an off-target homology.

3) Off-target sites cannot have indels, either DNA or RNA bulge. This assumption is against the existing biological discoveries, which will be discussed later.

In the proposed algorithm, the message recipient is the "bacteria" while the sender is the "phage". After some time of "evolution", the phage is aware of what sgRNA spacers the bacteria can recognize, a knowledge between the phage and bacteria only. For demonstration purpose, assume that the bacteria by far can recognize the following spacer sequence:

←←←

Position – 0 9 8 7 6 5 4 3 2 1 0 9 8 7 6 5 4 3 2 1

Spacer (S) – ACGTCGTAACGCGTATATGC

Using the same binary coding rule ((A:00), (C:01), (G:10), (T:11)), four nucleotide bases are required to define an eight-bit character. Thus, a 20-base sequence can code five characters. The substitution function is defined as:

1. T-A → 11
2. T-C, G-A → 10
3. T-G, G-C, C-A → 01
4. All other substitutions → 00

Note, T-A indicates that T substitutes for A. Suppose the secret message M is 01001110, let R be a large DNA sequence that has no homologous sequences of the spacer S (if there are, such homologous sequences must be either deleted or changed). To integrate M into R by substitution, the following algorithm is followed:

Step 1: Pick a large DNA sequence R, confirm there is at least one S inside R (must have PAM downstream). If there is none, substitute some nucleotides in R to generate S or select another valid R.

Step 2: compute the value of S by addition operation. The above sequence S = 30. Let v = (S mod 5) + 1 = 1. Only off-target sequences bearing exactly v mismatches with S would be used for carrying message. In this specific example, v = 1. Similarly, if v = 5, then valid off-target sequences must bear 5 mismatches.

Step 3: Change existing off-target sequences in R that bear exactly 1 mismatch with S so that they won't be mistaken as valid off-target sequences.

Step 4: Since M has eight bits, i.e. four nucleotides, four off-target sites are needed in R. Randomly but sequentially identify four non-overlapping sgRNA spacer sites in R, modify the site sequences according to the substitution function so that they become valid off-target sites of S. For instance, the message 01001110 can be fragmented into four off-target sites in order (PAM is attached):

Spacer: ACGTCGTAACGCGTATATGC

Site 1: ACGTCGTAACGCGTA**G**ATGC-GGG

Site 2: ACGTCGTAACG**G**GTATATGC-TGG

Site 3: ACG**A**CGTAACGCGTATATGC-CGG

Site 4: ACGTCG**C**AACGCGTATATGC-AGG

After the substitutions, carrier sequence R becomes R' and R' is sent to the recipient (bacterium) together with other noise DNA sequences (Note that these noise sequences do not contain S or any other bacteria-recognized CRISPR spacers). Once the bacterium receives R', it uses its existing CRISPR spacer sequence(s) to examine R'. If there is no recognized spacer in R', ignore it. In our case, as there is a recognized spacer S, the bacterium begins to process the R' sequence. The data retrieval reverses the protocol of data hiding.

IV. Results and Analysis

Unlike the classical substitution method, our method does not employ a one-to-one substitution function. We consider a one-to-one function is not necessary and one-to-one functions may present a more discernable pattern to intruders. The cracking probability of the above proposed method depends on the number of potential CRISPR spacers in the reference DNA sequence, the available substitution functions and binary coding rules. Since a spacer must be followed by an NGG PAM, the number of potential CRISPR spacers is about L/16, where L is the length of the reference sequence. The number of binary coding rules determines the available substitution functions, and it is $4 \times 3 \times 2 = 24$. Including the possible mismatch numbers, the cracking probability of the proposed method can be computed as:

$$\frac{16}{5 \times 24L} = \frac{2}{15L}$$

Thus, the cracking probability is a function of the size of the reference DNA sequence. This cracking probability is no better than that of the classical substitution method. To improve the cracking probability, we can remove the bacteria-recognized spacer S from R' because both the phage and bacteria are aware of the recognized spacers. In this revised scenario, to steal the secret message, the intruder must have knowledge of the spacer S. Therefore, the new cracking probability can be expressed as:

$$\frac{1}{120 \times 4^{20}} = 7.58 \times 10^{-15}$$

This cracking probability is much lower than that of the classical substitution method and it is independent of the reference DNA size. However, in computational practices, intruders can utilize sequence alignment techniques to guide their guessing directions and therefore greatly accelerate the cracking process. This is true to both our method and the classical method. To battle such a cracking strategy, the phage can make some random noise mismatches at other non-recognized CRISPR spacer locations.

In Table I, |M| = the length of message M in bits, L = the length of the reference DNA sequence, and *bpn* is the number of bits hidden per nucleotide. Note that our method doubles the *bpn* compared to the traditional substitution method. We also introduce a new concept "volume", which represents the maximum number of bits that the method can integrate into the reference DNA sequence. An off-target site must be of 23-nucleotides, and 5 mismatches can record 10 bits, thus the volume of our method is 10L/23, less than half of the classical method.

Allowing an indel in off-target sites can increase the cracking difficulty significantly as locating bulges is a much more time-consuming process than locating base mismatches. It has been proven biologically that valid off-target sites can have indels [25]. The reason why we didn't include indels is mostly for simplicity. A more sophisticated version can certainly include indels. For example, the 4th definition of the substitution function can be re-defined as: any a 1-nucleotide indel → 00.

We randomly generated 1000 sgRNA spacers, and then searched for their off-target sites in human genome. The result is summarized in Table II. The same computational experiment was conducted on a simulated genome of size 3 billion base pairs in which A, C, G, T are randomly distributed. The result is presented in Table III. The data in both Table II and Table III illustrate that the naturally-occurred off-target sites are not abundant, indicating noise mismatches must be added into the reference DNA sequence to distract the intruders. Otherwise, by sequence aligning the R and R', if R is publicly available, the intruders can quickly identify the spacer and retrieve the hidden message. Note that a simulated DNA sequence R can be created randomly anytime and can be deleted after R' has been generated because the recipient can retrieve the data from R' alone. Therefore, deleting R can disable the use of sequence alignment in cracking. Thus, it can be concluded that our proposed method works better with a simulated DNA sequence.

TABLE. I. PERFORMANCE OF OUR METHOD AND THE CLASSICAL METHOD

Method type	Capacity	Payload	bpn	Volume
Traditional	L	0	|M|/L	L
Our method	L	0	|M|/2L	10L/23

TABLE. II. OFF-TARGET SITES IN HUMAN GENOME

Number of mismatches	Total number of off-target sites	Averaged number of off-target sites per million bases
0	0	0
1	25	0
2	511	0
3	3306	0.001
4	45684	0.015
5	474587	0.158

TABLE. III. OFF-TARGET SITES IN SIMULATED GENOME

Number of mismatches	Total number of off-target sites	Averaged number of off-target sites per million bases
0	1	0
1	50	0
2	825	0
3	5376	0.002
4	67210	0.022
5	643482	0.214

The proposed method is not limited to digital DNA sequence applications. It can be practiced with biological DNA substances, too. One exemplar is to construct a plasmid and hide the secret message into the plasmid DNA sequence. The plasmid can be easily delivered to the recipient. The recipient applies Cas9-sgRNA reaction on the DNA. By analyzing the modified DNA sequences (for instance, deep sequencing), the recipient can detect all the off-target sites in the plasmid DNA sequence and from there, the recipient can then retrieve the hidden message. However, when practicing with biological DNA substances, one caution we must take is that, a large portion of potential off-target sites identified by computational methods do not exhibit any off-target effect in biological experiments [22, 26, 27]. Thus, only confirmed off-target sites should be included in the plasmid sequence.

The proposed method cannot be applied on DNA sequences of living organisms. The small number of naturally occurred off-target sites for a single sgRNA spacer limits the size of data that can be hidden in the reference DNA sequence.

V. CONCLUSION

In this paper, an original DNA based steganography method is proposed. This method adopts the CRISPR-Cas9 off-target editing and can reach much lower cracking probability than the classical substitution method. While the off-target editing is a flaw in CRISPR-Cas9 biological and medical applications, it can be used to enhance the applications of DNA based steganography.

REFERENCES

[1] K. Bailey and K. Curran, "An Evaluation of Image Based Steganography," Multimedia Tools and Applications, vol. 30, no. 1, pp. 55-88, 2006.

[2] C. T. Clelland, V. Risca and C. Bancroft, "Hiding messages in DNA microdots," Nature, vol. 399, pp. 533-534, 1999.

[3] A. Leier, C. Richter, W. Banzhaf and H. Rauhe, "Cryptography with DNA binary strands," BioSystems, vol. 57, pp. 13-22, 2000.

[4] H. J. Shiu, K. L. Ng, J. F. Fang, R. C. Lee and C. H. Huang, "Data hiding methods based upon DNA sequences," Information Sciences, vol. 180, pp. 2196-2208, 2010.

[5] L. M. Adleman, "Molecular computation of solutions to combinatorial problems," Science, vol. 266, pp. 1021-1024, 1994.

[6] D. Heider and A. Barnekow, "DNA-based watermarks using the DNA-Crypt algorithm," BMC Bioinformatics, vol. 8, p. 176, 2007.

[7] A. Khalifa and A. Atito, "High-capacity DNA-based steganography," in The 8th International Conference on Informatics and Systems, Giza, 2012.

[8] P. Das, S. Deb, N. Kar and B. Bhattacharya, "An improved DNA based dual cover steganography," Procedia Computer Science, vol. 46, pp. 604-611, 2015.

[9] P. Malathi, M. Manoaj, R. Manoj, V. Raghavan and R. E. Vinodhini, "Highly improved DNA based steganography," Procedia, vol. 115, pp. 651-659, 2017.

[10] D. A. Zebari, H. H. Haron and S. R. Zeebaree, "Security issuesin DNA based on data hiding: a review," International Journal of Applied Engineering Research, vol. 12, pp. 15363-15377, 2017.

[11] G. Hamed, M. Marey, S. E.-S. Amin and M. F. Tolba, "Hybrid, Randomized and high capacity conservative mutations DNA-based steganography for large sized data," Biosystems, vol. 167, pp. 47-61, 2018.

[12] H. Bae, B. Lee, S. Kwon and S. Yoon, "DNA steganalysis using deep recurrent neural networks," in Pacific Symposium on Biocomputing, Hawaii, 2019.

[13] S.-Y. Li, J.-K. Liu, G.-P. Zhao and J. Wang, "CADS: CRISPR/Cas12a-assisted DNA steganography for securing the storage and transfer of DNA-encoded information," ACS Synthetic Biology, vol. 7, pp. 1174-1178, 2018.

[14] NCBI, [Online]. Available: https://www.ncbi.nlm.nih.gov/genbank/statistics/. [Accessed 22 June 2019].

[15] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna and E. Charpentier, "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity," Science, vol. 337, pp. 816-821, 2012.

[16] R. Herai, "Avoiding the off-target effect of CRISPR/cas9 system is still a challenging accomplishment for genetic transformation," Gene, vol. 700, pp. 176-178, 2019.

[17] H. Y. Shin, C. Wang, H. K. Lee, K. H. Yoo, X. Zeng, T. Kuhns, C. M. Yang, T. Mohr, C. Liu and L. Hennighausen, "CRISPR/Cas9 targeting events cause complex deletions and insertions at 17 sites in the mouse genome," Nature Communications, vol. 8, p. 15464, 2017.

[18] D. Kim, D.-e. Kim, G. Lee, S.-l. Cho and J.-S. Kim, "Genome-wide target specificity of CRISPR RNA-guided adenine base editors," Nature Biotechnology, vol. 37, pp. 430-435, 2019.

[19] J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, H. W. Virgin, J. Listgarten and D. E. Root, "Optimized sgRNA design to maximize activity and minimize off-target effects of CRISCRISCRISCRISPR-Cas9," Nature Biotechnology, vol. 34, no. 2, pp. 184-191, 2016.

[20] H. Zhou, M. T. Zhou, D. Li, J. Manthey, E. Lioutikova, H. Wang and X. Zeng, "Whole genome analysis of CRISPR Cas9 sgRNA off-target homologies via an efficient computational algorithm," BMC Genomics, vol. 18, no. Suppl 9, p. 826, 2017.

[21] S. Bae, J. Park and J. S. Kim, "Cas-OFFinder: a fast and versatile algorithm that searches potential off-target sites of Cas9 RNA-guided endonuclease," Bioinformatics, vol. 30, pp. 1743-1745, 2014.

[22] P. D. Hsu, D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann, V. Agarwala, Y. Li, E. J. Fine, X. Wu, O. Shalem, T. J. Cradick, L. A. Marraffini, G. Bao and F. Zhang, "DNA targeting specificity of RNA-guided Cas9 nucleases," Nature Biotechnology, vol. 31, pp. 827-832, 2013.

[23] Y. Fu, J. D. Sander, D. Reyon, V. M. Cascio and J. K. Joung, "Improving CRISPR-Cas nuclease specificity using truncated guide RNAs," Nature Biotechnology, vol. 32, no. 3, pp. 279-284, 2014.

[24] D. Zhang, T. Hurst, D. Duand and S. J. Chen, "Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design," PNAS, vol. 116, no. 18, pp. 8693-8698, 2019.

[25] Y. Lin, T. J. Cradick, M. T. Brown, H. Deshmukh, P. Ranjan, N. Sarode, B. M. Wile, P. M. Vertino, F. J. Stewart and G. Bao, "CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences," Nucleic Acids Research, vol. 42, no. 11, p. 7473–7485, 2014.

[26] Y. Fu, J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, K. J. Joung and J. D. Sander, "High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells," Nature Biotechnology, vol. 31, no. 9, pp. 822-826, 2013.

[27] V. Pattanayak, S. Lin, J. P. Guilinger, E. Ma, J. A. Doudna and D. R. Liu, "High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity," Nature Biotechnology, vol. 31, no. 9, pp. 839-843, 2013.

A Literature Review on Medicine Recommender Systems

Benjamin Stark[1], Constanze Knahl[2], Mert Aydin[3], Karim Elish[4]
Department of Computer Science, Florida Polytechnic University, Lakeland, USA

***Abstract*—Medicine recommender systems can assist the medical care providers with the selection of an appropriate medication for the patients. The advanced technologies available nowadays can help developing such recommendation systems which can lead to more concise decisions. Many existing medicine recommendation systems are developed based on different algorithms. Thus, it is crucial to understand the state-of-the-art developments of these systems, their advantages and disadvantages as well as areas which require more research. In this paper, we conduct a literature review on the existing solutions for medicine recommender systems, describe and compare them based on various features, and present future research directions.**

Keywords—Medicine; recommendation systems; healthcare; systematic review

I. Introduction

Hospitals have access to vast amount of data about patients and their health parameters. Thus, there is a need for convenient way for medical professionals to utilize this information effectively. An example would be the access to aggregated information from existing database on a specific problem at the point of care when it is necessary. Moreover, there are more drugs, tests, and treatment recommendations (e.g. evidence-based medicine or clinical pathways) available for medical staff every day. Thus, it becomes increasingly difficult for them to decide which treatment to provide to a patient based on her symptoms, test results or previous medical history. On the other hand, all these data can be used to strive personalized healthcare which is currently on the rise and predicted to get a major disruptive trend in healthcare in the upcoming years.

Therefore, a recommendation engine for medical use could be employed to fill this gap and support decision making during therapy. Based on a patient's current health status, prehistory, current medications, symptoms and past treatments, the engine can look for individuals with similar parameters in the database. At the end, the recommender system would suggest the drugs that were most successful for similar patients. With the help of such a system, the doctor will be able to make a better-informed decision on how to treat a patient.

IBM's artificial intelligence machine Watson Health [8] is already able to find a suitable treatment for patients based on other patients' outcome and evidence-based medicine. IBM claims that 81% of healthcare executives familiar with Watson Health agreed that it has a positive impact on their business. This demonstrates that using technology and analytics become increasingly important in healthcare.

In this paper, we review the existing medicine recommendation system solutions, and compare them based on various features. The goal is to demonstrate the existing solutions for the healthcare providers in order to improve the medicine selection process and select an appropriate medication for the patients.

The rest of this paper is organized as follows: The methodology for the literature review is presented in Section 2. In Section 3, we discuss the findings. Section 4 presents the limitations. Finally, Section 5 concludes the paper and presents the future work.

II. Research Method

We conducted our literature review in several steps. We followed the guidelines defined in [9]. First, we defined search terms based on population, intervention, outcome of relevance and experimental design. However, we concluded that for our approach the population contains all healthcare facilities. Since this population is so comprehensive and non-specific, we excluded keywords about the population. This resulted in the following major keywords:

- Intervention: medication recommendation system
- Outcome of relevance: system for medication recommendation
- Experimental Design: empirical studies, systematic literature reviews, solution descriptions

The intervention and outcome of relevance category are the same. Therefore, they were only included one time. Once this has been agreed on, the search algorithm was constructed. The logical operators AND as well as OR were used to combine the search terms defined in the previous step. The following synonyms were considered:

- Medication: Medication, drug, Drug
- Recommendation: Recommendation, Recommender, recommender
- System: System, framework, Framework, algorithm, Algorithm, engine, Engine

This resulted in the following search algorithm:

{{medication OR Medication OR drug OR Drug} AND {recommendation OR Recommendation OR recommender OR Recommender} AND {system OR System OR Engine OR engine OR framework OR Framework OR algorithm OR Algorithm}}. To verify the algorithm and the terms used, we conducted a test for some papers we knew already. The test was successful as we could find relevant papers.

Afterwards, we chose the databases to search in based on available access which led to five databases:

- ACM Digital Library
- IEEExplore
- ScienceDirect
- Elsevier
- John Wiley Inc.

We also agreed on using Google Scholar as a search engine as a sixth source because it provides results from a high variety of databases which we might not have included and thus, can lead to a higher quantity of relevant papers.

Then, we agreed on inclusion and exclusion criteria which are defined as follows:

- Inclusion criteria:

1) Conference Proceedings and Journals published after 1999
2) Studies focusing on medicine recommendation systems in general and/or specified for any disease
3) Studies focusing on medicine recommendation systems based on graph databases

- Exclusion criteria:

1) Papers published before 2000
2) Manuscripts written in another language than English
3) Technical reports and white papers as well as Graduation projects, Master thesis and PhD dissertation
4) Textbooks (print and electronic)
5) Studies in other domains of knowledge

Finally, some quality criteria for the papers which met the inclusion criteria were defined to guarantee a selection of high-quality papers only. A scored system was used. For each of the following criteria met, a paper is assigned one point:

- Logical and reasonable in results and findings regarding the domain of knowledge
- Clearly stated objectives, results and findings regarding the domain of knowledge
- Well-presented and justified arguments
- Reasonably tested and/or applied system
- Well referenced with a minimum of ten sources

Only papers which met all criteria, thus had 5 points, were included in the final selection.

III. Results and Discussion

This section describes the final collection of papers in more detail, compares them with each other regarding different parameters, summarizes their approaches and defines research gaps. Table I presents the numbers of papers for the initial and final phase as well as the rate of included papers in percent. Also, to create more transparency, we included a column per phase for the results of the search with Google Scholar.

As shown in Table I, 52 documents were included initially. After the screening and cross-evaluation as described above, 13 documents remained. The IEEE database has the highest inclusion rate with 5 from the initial papers being included in the final set (71%). From the initial 22 papers from the ACM database 3 remained, leading to an inclusion rate of 14%. Also, it is surprising that none of the papers of ScienceDirect, Elsevier and John Wiley Inc. could be included in the final set of documents. One reason for that might be the publications about recommendation systems are not published in those databases, maybe because the editors of those publications prefer other journals and conferences.

Furthermore, for Google Scholar, there were 11 different publication venues included in the initial selection, from which 5 were included in the final selection. Five of the initial papers retrieved from Google Scholar (29%) were published in IEEE. In the final selection, 44% of the papers from Google Scholar were published by the IEEE. This leads to the conclusion that medicine recommendation systems are a widely discussed topic with many specifics which can also be recognized by the different areas of the journals, but the IEEE database seems to be the most attractive venue for publication in this area. However, in general, approaches and techniques related to medicine recommendation systems are published in a high variety of journals.

Specific journals, such as the Journal of Biomedical Semantics, seem to be promising for future literature research in this area, although the results retrieved via the IEEE database met by far the most the inclusion and quality criteria. Moreover, more than 50% of the documents from Google Scholar were included. This is not surprising since Google Scholar fetches documents from many different databases. Despite the strict quality criteria of our study, 25% of all initially selected papers were included in the final set of papers. This indicates a high quality of the databases searched in.

A. *Categorization of Approaches*

All the papers included in the final selection are categorized and summarized.

1) Ontology and rule-based medicine recommendation systems: The drug recommendation system GalenOWL [4] is based on the Greek drug guide GALINOS where doctors can search for a drug and find details on the drugs and additional information, such as interactions with other drugs. The paper describes a system that recommends drugs for a patient based on the disease of the patient, allergies and known drug interactions for the drugs in the database. To recommend the best fitting drug, rules for medications and interactions are stored in the system, which is based on ontologies, ICD-codes and other information. The application is accessible via the browser.

TABLE. I. NUMBER OF PAPERS FOUND IN THE RESPECTIVE SEARCH ENGINES (INCL. GOOGLE SCHOLAR RESULTS AND FINAL INCLUSION RATE)

Resource	**Initial Total (incl. Google Scholar)**	**Initial Google Scholar only**	**Final Total (incl. Google Scholar)**	**Final Google Scholar only**	**Inclusion rate in % Total (incl. Google Scholar)**
ACM	22	2	3	0	14%
IEEExplore	7	5	5	4	71%
ScienceDirect	5	0	0	0	0%
Elsevier	5	0	0	0	0%
John Wiley Inc.	3	0	0	0	0%
Springer	1	1	1	1	100%
Americana Medical Informatics Association	1	1	1	1	100%
Journal of Biomedical Semantics	2	2	2	2	100%
International Journal of Environmental Research and Public Health	1	1	0	0	0%
International Journal of e-Education, e-Business, e-Management and e-Learning	1	1	0	0	0%
Journal of Chemical and Pharmaceutical Sciences	1	1	0	0	0%
Advanced Internet of Things	1	1	1	1	100%
Clinical Pharmacology & Therapeutics	1	1	0	0	0%
International Journal of Computer Applications	1	1	0	0	0%
Total	52	17	13	9	25%

The drug-drug and drug-interaction discovery framework Panacea [5] is based on the approach GalenOWL and uses standardized medical terms and a rich knowledge base which are both modeled as rules. They used SKOS vocabulary, an ontology and reasoning engine and a medical and rules-based reasoning approach. The results show that Panacea is a promising solution, but still needs some improvement.

SemMed [14] which is a medical recommendation engine based on Semantic Web Technologies, applies an ontology-based approach. It consists of an inference engine, a rules manager, a support database, and ontology manager. The core classes "Diseases", "Medicines" and "Allergies" were used to develop rules.

Another solution proposed by [2] utilizes an ontology for anti-diabetic drug recommendations. However, it also includes the Multiple Criteria Decision-Making approach to compute weights and rank the drugs. It mainly utilizes laboratory data, but also considers risk and benefit factors.

IRS-T2D [11] is a drug recommender system which was specifically developed to individualize patient treatment of type 2 diabetes mellitus patients. The solution combines rule-based decision making with ontologies and semantic web technologies while taking specific patient information, such as the individual HbA1c target, into consideration.

Chen et al. [3] used semantic web rule language to describe the relationship between the rules retrieved from AACEMG. With the rules and knowledge from patient ontology and medicine ontology an inference is derived utilizing the Java Expert System Shell. The inference is then displayed in the system interface.

2) Data mining and machine learning-based medicine recommendation systems: The approach proposed by Sun et al. [15] analyzed EMR records to detect typical treatment regiments and measures (quantitatively) the effectiveness for those regimens for specific patient cohorts. The authors measure the similarity between the treatment records in the EMR, cluster similar ones to treatment regimens based on Map Reduce Enhanced Density Peaks based Clustering, extract semantically meaningful information for the doctor and estimate the treatment outcome for a patient cohort for a typical treatment regimen. The results of applying this approach in an empirical study show that the effective rate of the patient increases as well as the cure rate.

Hamed et al. [7] utilized Tweets from Twitter to analyze the well-being of the Tweeter and to give recommendations about alternative medicine possibilities. Therefore, the authors get the information of the Tweets, send the Tweeter a questionnaire to get more information about her state and apply a trained C4.5 decision tree algorithm to predict the condition of the user. Based on that, the algorithm can derive a recommendation for an alternative medical product.

DiaTrack was developed by Medvedeva et al. [12] as a drug recommendation system for type 2 diabetes and intends to give doctors a dashboard where they can see similar patient cases and their reaction to a drug or other factors. Therefore, the system compares the disease pattern of multiple patients and gives back the results in a color-coded, easy to understand graph.

The approach proposed by Kushwaha et al. [10] describes a drug recommendation system based on semantic web technology and data mining algorithms. Those two methods were combined to first extract semantic data and then apply data mining algorithms on those data. Data mining algorithms were used to individualize the treatment dependent on the patient's attributes. The system will not recommend drugs which the patient took before or that would interact with drugs the patient took before.

A hybrid framework to recommend drugs by ranking is proposed in [16]. Practitioners make inquiries and order lab tests. Information about this patient is entered into the system as a new case during the process. The system will process the new data and extract patient features. A diagnosis is made based on the patient's problem. The diagnosis is matched to a specific disease category in the system to determine which symptom-drug classifier to use. Patient features in the new case are put into the classifier to predict which drug cluster/clusters to choose for this patient. Drugs in each cluster will be ranked by the ranking module to form the final recommendation list.

Mahmoud et al. [1] investigated three different algorithms: Support Vector Machine (SVM), Back Propagation neural network, and ID3 decision tree to find out which algorithm is optimal for a drug recommendation framework. The evaluation is based on scalability, accuracy, and efficiency. Since accuracy is the most important criteria for recommending a drug, the SVM algorithm was identified as the most useful algorithm. The next steps are to implement the model along with the data preparation, visualization, and database system module.

Another drug recommendation system is a cloud-based platform utilizing various algorithms [17]. Using the vector service model, the drug character is formatted according to the description of the drug information. Then a k-means algorithm is applied to cluster drugs. Subsequently, an evaluation using collaborative filtering leads to recommendations. Finally, tensor decomposition is applied to address sparsity and massive data, shortcomings of collaborative filtering. This multi-step process helps to make an accurate recommendation.

B. Characteristics of Approaches

Tables II, III and IV present the results of the literature review. The columns refer to the different dimensions we compare the studies with. In each table, there is a short summary and discussion of each column presented in the corresponding table.

In Table II, the column "Disease" describes whether the concept described in the study focuses on a specific disease. "Data storage" summarizes the method applied to store the data. "Interface" refers to the connection of the back-end modules with each other and the front-end. The column "Data collection" describes the sources the data used for testing the approach, if applicable, were gathered from.

1) Disease: As shown in Table II, most studies do not focus on a particular disease. This shows that most work in this field attempts to develop a general-purpose recommendation engine. Finding a recommender system that will work for all diseases would be very useful for general practitioners. However, all studies dealing with a specific disease focus on drug recommendation for diabetes. This means that this type of disease seems to be relatively important and well suited for a drug recommendation system. Since a highly-individualized treatment is required for diabetes, this is also reasonable.

2) Data storage: Data storage is not widely discussed in the studies we reviewed. 5 out of 13 papers mention their data storage approach for datasets such as patient data and drug data. This shows that mostly it is preferred to focus on selected parts of the solution, such as the algorithm. For the studies that include data storage, they all have different ways to store data sets.

GalenOWL [4] stores data in RDF graphs and utilizes SPARQL queries whereas the SWRL [3] leverages a software called Protégé [6] to store its data. Author in [17] utilizes cloud storage services and the IRS-T2D [11] applies ontologies and semantic web technologies. This shows that there is no standardized approach to store data although the data sets comprise similar data from the electronic medical record (EMR).

3) Interface: Little information is provided about the interface of drug recommendation systems. The focus is on the recommendation algorithm. Two studies utilize Protégé for the interface and semantic web rule language to show the output of the result of the algorithm. On the other hand, DiaTrack [12] leverages dynamic-service middleware to provide a visualization of the output. This shows that drug recommendations are still in development. Generally, once the recommendation engine is defined, it seems like the focus is on the user experience. Moreover, for the studies that did provide information about user interfaces, Protégé is the framework mostly applied. Protégé [13] is an open-source ontology editor that provides developers with a user interface to create intelligent systems. Developed by Stanford University, this application is appealing due to its free-to-use license terms, its active community for support and its extensible environment.

TABLE. II. RESULTS OF LITERATURE REVIEW IN TERMS OF DISEASE, DATA STORAGE, INTERFACE AND DATA COLLECTION USED IN THE STUDIES

Study	**Disease**	**Data storage**	**Interface**	**Data collection**
Data-driven Automatic Treatment Regimen Development and Recommendation [15]	Not specified	No information	No information	14 grade 3 hospitals in China
Panacea, a semantic-enabled drug recommendations discovery framework [5]	Not specified	No information	No information	University hospital of Thessaloniki
SemMed [14]	Not specified	No information	No information	No information
The recommendation of medicines based on multiple criteria decision making and domain ontology [2]	Diabetes (not further specified)	No information	No information	No information
T-Recs [7]	Not specified (Author tested it for symptoms such as diarrhea, headache, fever, stomach ache)	No information	No information	Twitter (~500,000 Tweets)
GalenOWL [4]	Not specified	RDF graph SPARQL queries	No information	No information
IRS-T2D [11]	Type 2 Diabetes	Knowledge base with ontologies and SWR	Jess SWRL2Jess OWL2Jess Protégé-OWL API	Test patients
DiaTrack [12]	Type 2 Diabetes	Database management system	Dynamic service middle-ware (similarity, reasoning and visualization components and provides professional interaction tools)	No information
LOD Cloud Mining for Prognosis Model [10]	Not specified	No information	No information	Various Semantic Web sources (SIDER (for side effects), Drug Bank)
A framework of hybrid recommender system [16]	Not specified	No information	No information	Electronic Medical Record System (EMRS)
A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection [3]	Type 2 Diabetes	Protégé	Protégé	No information
An Intelligent Medicine Recommender System Framework [1]	Not specified	No information	No information	Hospital Information System (HIS) with 70% training data and 30% testing data.
CADRE [17]	Not specified	Cloud	No information	Walgreens

4) Data collection: Generally, patient data are collected from hospital information systems (HIS). Data retrieved from HIS accurately resemble the data hospitals have available in practice. HIS data are usually not in an appropriate format for most ontology-based intelligent systems. Hence, it is essential to think of ways to format the data so they fit to the algorithm applied. For the drug data, various sources were leveraged. CADRE [17] used the Walgreens website whereas [10] include data from a drug bank and SIDER for side effects. Comprising different sources for drug data could lead to discrepancies in the description of benefits as well as in the side effects of drugs. These differences might influence the drug recommendation from the data perspective. Furthermore, data are in different languages. For example, the data used by [15] include data in the Chinese language, since they are retrieved from Chinese hospitals. Thus, it is important to assess the semantics when dealing with multiple languages. Furthermore, sentiment data are generally not collected except for T-Recs [7], which used 500,000 tweets from Twitter. This means that the proposed recommendation systems generally do not consider patient feedback. This is an important element that is being omitted and might reduce overall patient satisfaction.

In Table III, the column “Data preparation” relates to the steps applied to the raw data so they fit for the algorithm. “Platform/Technology” refers to the technology and/or platform utilized for the implementation.

5) Data preparation: In Table III, we found that the studies apply different data preparation procedures, for instance regarding data formatting and cleaning. This seems reasonable if we assume that the format of data is different across the different studies. Furthermore, the table shows that most studies utilize data preparation modules to provide an acceptable format for their algorithm module. Studies such as [1], [16] or [2] use normalization techniques to uniformly scale the data across the modules. On the other hand, some authors decided to manually prepare the data. In the case of T-Recs [7], tweets were distinguished to be either relevant or irrelevant. The study in [15] shows the most relevant medicines were selected and then divided into four periods. Since most studies have information about data preparation, it shows that this aspect is essential when developing a recommendation engine.

6) Platform/Technology: With regards to the platform/technology of the recommender systems, three of the studies use online services. CADRE [17] utilizes the cloud platform to give medicine recommendations based on symptoms. The LOD cloud mining study by [10] leverages semantic knowledge from the LOD cloud. The GalenOWL [4] uses semantic-enabled online services to provide drug-drug and drug-disease interaction discovery. Furthermore, T-Recs [7] utilizes Twitter to monitor tweet sentiment, create an analysis for the tweets, and the calculate recommendations. Other studies apply rule-based inference engines such as Pellet and Jena/Drools. In all 13 studies, the technology used to apply the algorithm was different. This shows that researchers do not restrict themselves to apply only one specific software tool, but utilize the various possibilities available. Despite this flexibility, scientists need to consider the costs of the technology to make it reasonable for an average hospital to purchase it. Hence, open-source software which was used to develop algorithms such as Protégé, Pellet and Jena rule engine seem to be reasonable and preferable choice.

Table IV compares the studies in terms of the algorithms used, and presents future work identified by these studies. The algorithms used in the reviewed studies were described earlier in Section III.

TABLE. III. RESULTS OF LITERATURE REVIEW IN TERMS OF DATA PREPARATION AND THE PLATFORM/TECHNOLOGY USED IN THE STUDIES

Study	**Data preparation**	**Platform/Technology**
Data-driven Automatic Treatment Regimen Development and Recommendation [15]	Yes (select most relevant medicines (138); divide treatment record into 4 periods)	Custom
Panacea, a semantic-enabled drug recommendations discovery framework [5]	Yes (applying the SKOS vocabulary)	Querying instance and knowledge base Rule engines (Jena/Drools rule engine)
SemMed [14]	No information	Inference engine Rules manager Support DB and ontology manager
The recommendation of medicines based on multiple criteria decision making and domain ontology [2]	Yes (normalization of benefit and risk factors)	No information
T-Recs [7]	Yes (Manual distinction between relevant and irrelevant tweets; grouping of tweets)	Twitter Tweet Sentiment monitor Tweet analysis and computing recommendation
GalenOWL [4]	Yes (ATC, ICD-10, UNII, Substances, Conditions, Indications-Contraindications)	Online-service
IRS-T2D [11]	No information	No information
DiaTrack [12]	No information	A standard web-browser front-end for Data Entry, Research, Practice Administration and Site Administration
LOD Cloud Mining for Prognosis Model [10]	Yes (Queried with SPARQL)	LODD cloud, queries on drug data with SPARQL 1.1 with Java IDE, database: RDF dump stored in Sesame, app uses the server of Sesame
A framework of hybrid recommender system [16]	Yes (Text Mining Module, Data Normalization, Drug Clustering Module)	No information
A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection [3]	Yes (Inference engine (Pellet) transformed the format acceptable to the recommendation system)	Medicine ontology was created by Protégé Inference engine (Pellet)
An Intelligent Medicine Recommender System Framework [1]	Yes (Data normalization using min and max functions. Correlation analysis using Chi-Square Tests.)	No information
CADRE [17]	No information	Cloud

TABLE. IV. RESULTS OF LITERATURE REVIEW IN TERMS OF ALGORITHMS AND FUTURE WORK PRESENTED IN EACH STUDY

Study	Algorithm	Future work
Data-driven Automatic Treatment Regimen Development and Recommendation [15]	• Custom (not pre-defined, but used Map Reduce Enhanced Density Peaks based Clustering and decision tree)	• No information
Panacea, a semantic-enabled drug recommendations discovery framework [5]	• Ontology • Rules	• Weighing of interactions, contraindications based on severity observation, probabilistic inference • Addition of dosage recommendation • Increase sample for testing. • Automate manual work which is required to enrich rule base.
SemMed [14]	• Ontology • Semantic web techniques	• Integration of this system with other existing ones, e.g. MEDBOLI. • Addition of functions, e.g. prediction of treatment times with neural networks.
The recommendation of medicines based on multiple criteria decision making and domain ontology [2]	• Ontology (created with protégé) • Multiple Criteria Decision Making	• Evaluation by multiple diabetes physicians
GoT-Recs [7]	• Decision tree (C4.5)	• Extend it to other products, e.g. beverages, snacks, etc. • Include other online data as well, e.g. "PatientsLikeMe" • Get permission from twitter • Modify architecture so it is scalable • Test different algorithms
GalenOWL [4]	• Ontologies • Semantic web technologies • Rules	• Expansion of semantic rules • Prioritization of interactions of drugs and diseases since not all interactions have the same importance • Performance optimization (e.g. context extraction from medical knowledge)
IRS-T2D [11]	• Ontologies • Semantic web technologies	• Test more patients • Improve patient profile to store more information • Add insulin data and rules considering insulin to the system
DiaTrack [12]	• Pattern comparison in data	• Enhance system to not only work based previous patients but also based on general drug features (consider ontologies and semantic web technologies)
LOD Cloud Mining for Prognosis Model [10]	• Semantic Web techniques • Data mining algorithm (Decision trees usin • C4.5 algorithm and bagging)	• Update data on drugs, diseases and interactions as needed • Extract more meaningful features like toxicity, food interaction etc.
A framework of hybrid recommender system [16]	• Case-based reasoning for ranking the drug clusters • Artificial neural network for Symptom-Drug Classifier module • cTAKE system for text-mining module	• Implementation of the recommender system • Free text messages in EMRS need to be taken into consideration • Distinguishable features for personalization need to be extracted • The recommender must be dynamic and adaptive to assess temporal efficiency of drugs and add new drugs
A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection [3]	• Semantic Web Rule Language(SWRL) • Java Expert System Shell (JESS) • Ontology	• Strengthen patient ontology • Test more patient data for the system • Calculate the dosage of the medicine
An Intelligent Medicine Recommender System Framework [1]	• SVM (Support Vector Machine) • Back-propagation neural network • ID3 decision tree	• Build the recommendation model • Apply MapReduce to enlarge the ability of processing big diagnosis data
CADRE [17]	• Vector space model (VSM) • K-means clustering • Collaborative Filtering • Tensor decomposition	• Investigate how to improve the accuracy of CADRE by considering user's age, geography, and other factors

As the table indicates, almost all of the approaches (9 out of 13) state some future work and research areas. Although some of them are rather specific to the technique presented in the paper, it is possible to derive some general fields where more research is required. The three main areas based on this literature review are:

- Finding solutions for including a recommendation for the dosage of a medicine
- Verifying the results, e.g. by increasing testing, especially the sample of testing
- Finding solutions which are highly scalable

This extensive literature review shows that there are many solutions for drug recommendation systems. Most of them are based on manually constructed ontologies and use sophisticated data mining or machine learning methods. Especially the processes including manual work are very time consuming. Also, none of these approaches utilizes a graph database to model the relationships between patients and to apply an algorithm to this model, although this might be a well-suited approach. Graph databases can model the data in graphs which is a more natural way to store data than any other database offers. Medical institutions usually have many patients who can be illustrated in a graph as a network of patients. Thus, this approach may be superior to the ones discussed earlier in this paper and also addresses the last topic listed for future research. The reasons for that are the unique features of graph databases, such as high consistency and high scalability.

IV. Limitations

Our literature review has two main limitations, namely, the paper selection and content. Out of 52 papers, only 13 were reviewed based on the strict inclusion, exclusion and quality criteria we chose. Along with the strict search criteria, the systematic review included papers from a limited number of databases. However, we used six main databases that are well known.

Some papers offer little detail on the exact implementation and architecture of the solutions built. This made it more difficult to assess which applications were used to build the system. Also, some papers proposed only a theoretical solution on how to recommend a drug such as [16], but did not implement the solution. On the other hand, some papers did implement the solution such as [2], but no evaluation was made on the performance. Therefore, several questions stay under investigation, such as "how accurate are these recommender systems?" and "does it reduce the symptoms patients have?"

V. Conclusions and Future Work

This paper presented a systematic literature review for medicine recommendation engines. We reviewed 13 studies that met our strict criteria in six different databases. These studies can be split into two categories: (*i*) machine learning and data mining-based, and (*ii*) ontology and rule-based approach. The studies were summarized and evaluated across several parameters: diseases, data storage, interface, data collection, data preparation, platform/technology, algorithm, and future work. Most of the studies that did not focus on any disease, had less information about data storage, interface, data collection, data preparation, platforms and technology, and customized algorithms.

For future work, our review suggests to extend the existing solutions by adding recommendations for the dosage of drugs, as well as building highly scalable solutions. Also, based on the evaluation, we identified that none of the studies we reviewed include a graph database in their solution for a drug recommendation system. Graph database such as Neo4j seem to be very suitable for drug recommendation engines because they are highly scalable and consistent which would account for the last of the aforementioned topics for future work. Furthermore, their data model seems to be promising for recommendation systems due to their network structure and ease for querying. Hence, another direction for future research would be the creation of medicine recommendation engines based on graph database.

References

[1] Bao, Y. and Jiang, X. 2016. An Intelligent Medicine Recommender System Framework. 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA).

[2] Chen, R.-C., Chiu, J. Y., and Batj, C. T. 2011. The recommendation of medicines based on multiple criteria decision making and domain ontology — An example of anti-diabetic medicines, 27–32.

[3] Chen, R.-C., Huang, Y.-H., Bau, C.-T., and Chen, S.-M. 2012. A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. Expert Systems with Applications 39, 4, 3995–4006.

[4] Doulaverakis, C., Nikolaidis, G., Kleontas, A., and Kompatsiaris, I. 2012. GalenOWL: Ontology-based drug recommendations discovery. Journal of Biomedical Semantics 3, 14.

[5] Doulaverakis, C., Nikolaidis, G., Kleontas, A., and Kompatsiaris, I. 2014. Panacea, a semantic-enabled drug recommendations discovery framework. Journal of Biomedical Semantics 5, 13.

[6] Guo, W. Y. 2008. Reasoning with semantic web technologies in ubiquitous computing environment. Journal of Software 3, 8, 27–33.

[7] Hamed, A. A., Roose, R., Branicki, M., and Rubin, A. 2012. T-Recs: Time-aware Twitter-based Drug Recommender System. 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[8] IBM. 2017. IBM Watson Health. http://www.ibm.com/watson/health/.

[9] Kitchenham, B. and Charters, C. 2007. Guidelines for Performing Systematic Literature Reviews in Software Engineering. 2007 Joint Report - EBSE 20 07-0 01.

[10] Kushwaha, N., Goyal, R., Goel, P., Singla, S., and Vyas, O. P. 2014. LOD Cloud Mining for Prognosis Model(Case Study. Native App for Drug Recommender System). AIT 04, 03, 20–28.

[11] Mahmoud, N. and Elbeh, H. 2016. IRS-T2D. Individualize Recommendation System for Type2 Diabetes Medication Based on Ontology and SWRL. In Proceedings of the 10th International Conference on Informatics and Systems - INFOS '16. ACM Press, New York, New York, USA, 203–209. DOI=10.1145/2908446.2908495.

[12] Medvedeva, O., Knox, T., and Paul, J. 2007. DiaTrack. Web-based application for assisted decision-making in treatment of diabetes. Journal of Computing Sciences in Colleges 23, 1, 154–161.

[13] Protégé. 2016. Protégé. http://protege.stanford.edu/. Accessed 15 March 2017.

[14] Rodríguez, A., Jiménez, E., Fernández, J., Eccius, M., Gómez, J. M., Alor-Hernandez, G., Posada-Gomez, R., and Laufer, C. 2009. SemMed: Applying Semantic Web to Medical Recommendation Systems. 2009 First International Conference on Intensive Applications and Services.

[15] Sun, L., Liu, C., Guo, C., Xiong, H., and Xie, Y. 2016. Data-driven Automatic Treatment Regimen Development and Recommendation. KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1865–1874.

[16] Zhang, Q., Zhang, G., Lu, J., and Wu, D. 2015. A framework of hybrid recommender system for personalized clinical prescription. 2015 International Conference on Intelligent Systems and Knowledge Engineering.

[17] Zhang, Y., Zhang, D., Hassan, M. M., Alamri, A., and Peng, L. 2015. CADRE. Cloud-Assisted Drug REcommendation Service for Online Pharmacies. Mobile Netw Appl 20, 3, 348–355.

Study and Analysis of Delay Sensitive and Energy Efficient Routing Approach

Babar Ali[*,1], Tariq Mahmood[2], Muhammad Ayzed Mirza[3], Saleemullah Memon[4]
Muhammad Rashid[5], Ekang Francis Ajebesone[6]
School of Information & Communication Engineering
Beijing University of Post and Telecommunication, Beijing 100876, China[1, 4, 5, 6]
Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China[2]
Division of Science and Technology, The University of Education, Lahore 54000, Pakistan[2]
School of Electronic Engineering, Beijing University of Post and Telecommunication, Beijing 100876, China[3]

***Abstract*—Wireless Sensing Networks (WSNs) comprised of significant numbers of miniatures and reasonable sensor nodes, which sense data from surrounding and forwarded data toward the base station (BS) via multi-hop fashion through cluster head node (CHN). The random selection of CHN in WSNs is fully based on the nodes residing energy. The node residing energy and network sustainability is hot research issues of the day in WSNs. There are many deficiencies in less energy adaptive clustering hierarchy (LEACH) RP due to the rapid energy usage of ordinary and CHN because of direct communication to the base station. The rapid draining of node energy produces huge numbers of hole in the network causing retransmission of data packet, route update cost, and E2E delay. In this paper, the proposed Delay Sensitive and Energy Efficient (DSEE) Routing Protocol (RP) select CHN considering distance difference and amount of remaining energy of neighboring nodes. In this proposed approach, data fusion technology (DFT) was implemented to solve the problem of data redundancy, but it does not design a specific data fusion algorithm. At last, simulation experiments proved the superiority of the improved protocol LEACH-DSEE and finally, we compare this improved routing protocol with existing protocols by consideration metrics such as node death ratio, data packet delivery ration and node energy consumption.**

Keywords*—*Multi-hop (MH); CHN; WSNs; BS; Data Fusion Technology (DFT); LEACH RP; DSEE

I. Introduction

Nowadays efficient routing in WSNs is the hottest challenges of research. WSNs are characterized by the distribution of many miniature and inexpensive nodes in the monitoring area [1, 41]. WSNs is a collection of static sensing nodes that collects information and send to the BS. BS accumulate data sensed from the special discovering areas by using the main function of WSNs [2, 3]. WSNs are broadly used such that video surveillance, in which the sensor nodes(s-nodes) work as cameras that send the collected image information or video information to the destination. In the defense field, s-nodes are distributed in enemy areas that transmitted the enemy's related information to the command headquarters [4-6]. The important characteristics of s-nodes are uses of low power consumption for the short transmission range. The key point of the network data collection is, how to convey the collected information from the source nodes to rely on the node and then send to the BS. Nodes mainly undertake the task of collecting data and forwarding routing data. Routing protocol plays a key for the communication of information from the source node to BS. The important work of routing protocols is to select the optimal route for the node for data forwarding toward the base node [7]. In WSNs routing protocol is the primary technology that works under the network layer [8]. The routing protocol of the network layer ensures that s-node has successfully transmitted data toward the base station. It is very important to diminish the consumption of energy and encompass network life. In addition, WSNs are often large-scale application scenarios so design a routing association that can be applied in very long -scale WSNs. Finally, applying data fusion technology in routing protocol greatly reduce the data traffic to retain valid data resulted from a reduction in data redundancy. LEACH-RP solves the problems of a planar RP such as more delay, high energy usage [9, 10]. The practical application proves that LEACH RP has a 15% longer network lifetime than general planar RP [11]. However, LEACH RP still has some shortcomings like as uneven division and per round unstable numbering of the CHN, lack of consideration of residual energy during the assortment of CHN, etc. These shortcomings will increase the energy usage of nodes, also reduce the network period [12]. LEACH RP stipulates that all nodes and the BS communicate directly, so LEACH RP is not a suitable protocol for very large-scale WSNs application [13]. The working of LEACH RP is based on data fusion technology, but its model does not build on specific data fusion algorithm. Aiming at the shortcomings regarding LEACH RP design, we propose a LEACH-DSEE RP which can reduce the consumption of the energy and simulates this improved protocol. All experimental results prove that the LEACH-DSEE RP can proficiently reduce energy consumption as compared with different LEACH, LEACH-M LEACH-C, there have been many improvements based on original LEACH protocol in recent ten years. The most famous improvement RP is LEACH-C RP [14], the main idea of LEACH-C and LEACH-Mobile (M) RP is the choice of the best CHN set through BS control. For the performance measurement of lossy network & less power RP algorithm, various mobility models used such as GMM, RWP, MGM [15]. Two important issues in WSNs, as they can directly impact the lifetime of network & operation, one is sensing coverage and second is the connectivity of network [39, 40].

[*] Corresponding Author

II. Related Work

Fig. 1 explained the architecture of WSNs. WSNs comprises of many s-nodes, BS, Internet networks and servers. S-nodes are composed of four modules. The function of the sensor module is to collect data for monitoring area; the function of the processor module is to process collected data information, such as data fusion; communication module is responsible for communication; power module is responsible for providing electricity. WSNs establish data communication links from source nodes to destination nodes through self-organization [16-19]. In Fig. 1, five labeled nodes (ABCDE) form a data transmission path from the source node (Node A) to the final node (BS). BS forwards collected data to the offshore network station. WSNs is dissimilar from the traditional WSNs. There are many data collection protocols in WSNs [20-22]. LEACH is a low-energy adaptive layered routing protocol. Later, many-layered RP has enhanced behalf on the LEACH RP. The operation of LEACH RP [23-25] is periodic. Flooding protocol is a classic planar RP for cluster building and data collection. Its advantage is that it does not need to maintain routing information. All collected data from the source nodes are forwards to all neighbor nodes and the neighbor node forwards its own information and collected information to the next coming nodes until the end node receives the data. The PED&P (Power Efficient Data Gathering & Aggregation) in WSNs protocol systematizes the nodes in the whole system into a least spanning tree based on the BS which diminishes the consumption of energy in the whole system [26, 27].

LEACH-C RP [28] is a very famous LEACH improvement protocol. At the beginning of each round, it is controlled by the BS, which choose the optimal CHN set depending on the node's residual energy and location information. LEACH-C is different from the LEACH RP in clustering stage. The nodes of LEACH-C RP have the information about their location, all other non-CHN calculate their nearest CHN behalf on the location information of the CHN broadcasted by the BS and join its cluster. Literature [29] proposes an enhanced protocol based on LEACH protocol. The improved protocol also runs repeatedly in the unit of "wheel". One round is alienated into cluster building and data the collection stages. The improved protocol improves the threshold T_n formula.

$$T_n = \begin{cases} \frac{\text{Probability}(p)}{1-p\times[\text{rmod}\left(\frac{1}{n}\right)]} \times (\text{mE}_p(i) + \text{nN}(i)) \\ 0 \end{cases} \tag{1}$$

Ep (i) is nodes energy factor and If node I have a large amount of residual energy, Ep(i) is also large. The density factor of node N (i) and m, n are weight parameters, and 0 < m, n< 1, and m+ n = 1. Document [29] also improves the threshold Tn formula. The new threshold Tn considers the node residual energy factor & node distance factor from BS.

$$T_n = \begin{cases} \frac{\text{Nodes probability}(p_n)}{1-p_n\times\left[\text{r mod}\left(\frac{1}{p_n}\right)\right]} \times \left(1 + m \times k \times \left(\frac{E_{\text{current}}}{E_{\text{total}}}\right)\right) \\ +n \times d_2^2/(d_1^4 + d_2^2) \\ 0 \end{cases} \tag{2}$$

In equation (2), m and n are weight parameters and m + n =1. K is the expected number of CHN, residual energy $Ec_{\text{current}}/E_{\text{total}}$ ratio is b/w the current and all nodes, d_2 is the distance b/w the monitoring area center and node, d_1 is the distance b/w the node & BS. The improved protocol proposed in reference [30] deliberates the residual energy & distance from the BS during the selection of CH nodes. In literature [31, 32], new sort of hierarchical protocol was anticipated behalf on LEACH-C & original LEACH protocols. The proposed protocol attains energy proficiency through variation cluster heads selection formula which has huge energy and contribution of short role in the selection of CH / VCH and the steady-state (SS) phase. The introduction of VCH introduce the frequency of re-clustering poorer and extends the lifetime in the SS phase. The fertile way to make WSN more energy efficient is to split the network into the various cluster. In clustered networks, the ordinary node senses information from the environment and transmit it to CHN. The CHN gathering the sense data and perform the aggregation on collected data before sending to the BS to achieve the scalability, load balancing, data fusion, and reduction of e2e delay. The literation [33-35] explain the detailed survey of clustering techniques along with their characteristics and clustering algorithms. Literature [36-39] represents a solar CHN selection based on Solar-aware LEACH (sLEACH). The sLEACH protocol requires some sensor nodes that can convert solar energy into electrical energy and can increase their residual energy through the sun's irradiation. In the sLEACH protocol, Probability of solar node becoming CHN is increased and the life of the network is prolonged by solar energy. Because the sLEACH protocol adds solar energy to sensor nodes, it is more effective than the generally improved protocol in extending network life duration.

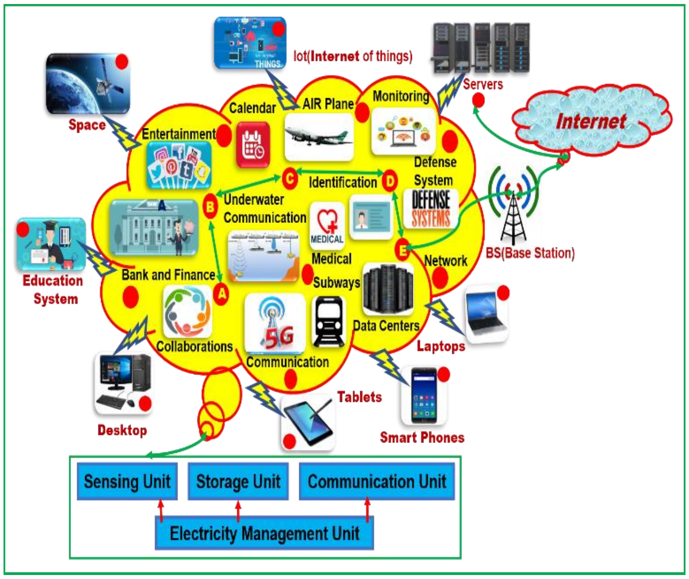


Fig. 1. The Architecture of WSNs.

III. Problem Description

The key inefficiencies of LEACH RP are the following:

- During the selection of CHN, LEACH RP does not deliberate the residual energy. If the amount of residual energy is relatively small, nodes acting as CHN & can represent to some premature dead nodes.
- LEACH RP does not require the number and distribution of CHN per round, which may cause clusters are unstable and the distribution is not uniform that resulting in maximal and minimal clusters.
- LEACH RP does not deliberate the distance b/w the CHN & BS. The longer the communication distance between sensor nodes, the more energy usage. So, the nodes far away from the BS may die prematurely.
- LEACH RP requires all CHN to communicate directly with the BS instead of the communication between sensor nodes. If the transmission distance is long, the data information cannot be sent and consume a lot of energy. Therefore, the LEACH RP is not appropriate for huge-range networks.
- LEACH RP requires a CHN to fuse data collected by cluster, but it is not designed specific data fusion algorithm.

IV. Architecture of Proposed Protocol LEACH-DSEE

This proposed approach mainly related to the reactive networks according to the needs of time depending on applications. It describes the problem of heavy energy usage during data forwarding and selection of CH Nodes. LEACH DSEE RP provides stable energy consumption by using direct communication between the CHN and BS.

A. Regional Clustering

The BS logically distributes the monitoring area into numerous small areas, each of which merges clusters according to the strength of nodes involved. If the number of sensor nodes in a small area is not less than threshold M, the small area is a cluster. If the number of nodes in a small area is less than threshold M, it merges with the surrounding small area until the number of nodes in the merged area is less than M, and the merged area becomes a cluster. Fig. 2 explained, all s-nodes are deployed randomly and uniformly in the experimental area. The matrix box is representing the area to be detected and the black dot is the s-node.

The specific number of small areas is determined according to the actual situation. The proposed scheme BS takes nine small areas that label each small area seen in Fig. 3. Finally, the BS divides clusters according to the number of nodes contained in each small area. BS receives the information consist of the node’s location of all sensor nodes that calculates the strength of nodes contained in each small area. If the merged area is less than threshold M, it continues to be merging the surrounding area till not less than the value of threshold M. In the proposed techniques, threshold M set on half of the average value.

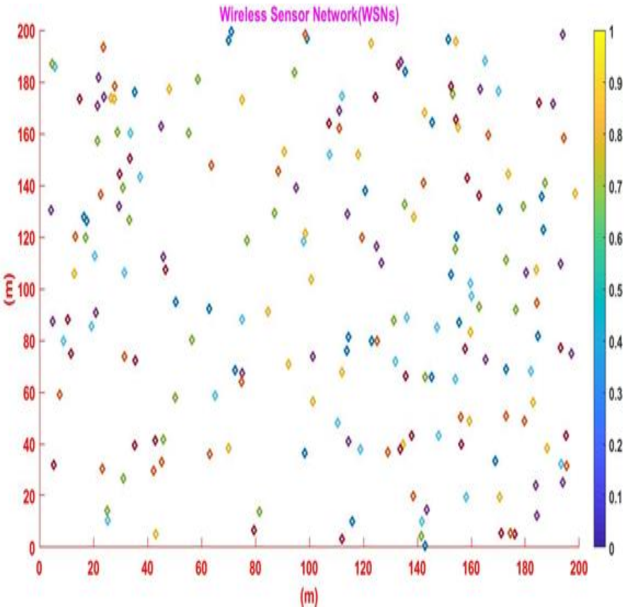


Fig. 2. Deployment of s-Nodes in Monitored Area.

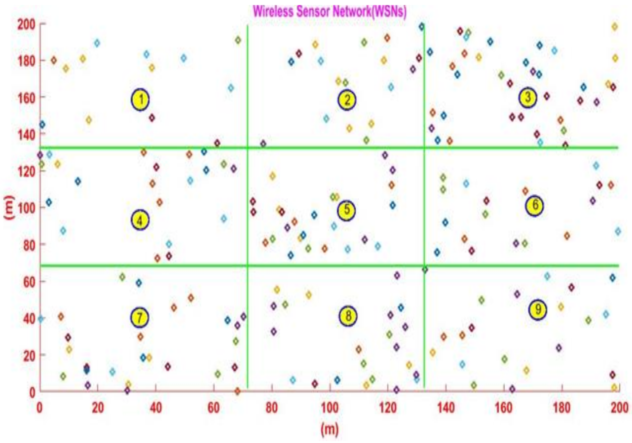


Fig. 3. Division of Monitoring Area.

B. Cluster Head Selection (CHS)

BS divides the monitoring area into several clusters according to the idea of regional clustering. At the starting of every round, every cluster chooses the CHN according to the largest amount of residual energy. We see this CHS in Fig. 4.

C. Inter-Cluster Transmission (ICT)

The proposed LEACH-DSEE routing protocol, solve the energy consumption issues of network nodes. CHN and BS transmit data through multi-hop and single hop fashion during communication. The nodes are alienated into three categories like nearest node, next nearest nodes, and a farther node according to the distance difference b/w the sensor and BS. The nearest CHN communicates directly to the destination/BS and also work as a relay node. The distant CHN also transmits data directly to the destination/BS and did not work as a relay node. Table I elaborate on the data structures of Cluster-head-node and Cluster Node of LEACH-DSEE.

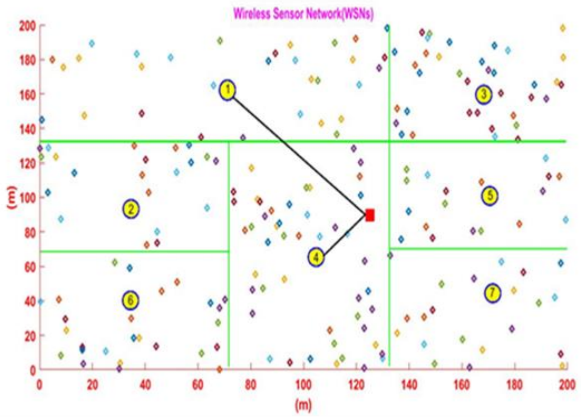


Fig. 4. Cluster Diagram of the Monitored Area.

TABLE I. ELEMENTS OF CHN AND CN

Name of Elements of CHN and CN	Initial Value
Node ID Number	Each Node has a Unique ID
Cluster Node	0
Residual Energy	1
Dead Node	0
Location of CHN Position	Each Node has its own Location
Cluster Number	0
CN Classification Based on Distance from BS	0

D. Re-Selection of CHN in Clusters

In the data collection phase, all cluster members send data information to CHN along with own residual energy and ID information. CHN retain the residual energy information of all its own cluster member nodes. After data collection, CHN utilizes a large amount of energy and Cluster-head-node need to select new CHN. The CHN broadcasts the ID information of the node having a large amount of residual energy in the cluster for the choice of new CHN. All cluster members send the "join cluster" message to the new CHN. Then the new CHN allocates channels to all cluster members and then continues to receive data collection stage.

V. ALGORITHM FOR THE SELECTION OF OPTIMAL FORWARDED NODES

Fig. 5 elaborate the routing decision based on residual energy. The CHN which is close to the BS sends the data directly to the BS and acts as the relay node. The distant CHN also communicate directly to the destination and did not work as a relay node.

VI. PROTOCOL SIMULATION AND VERIFICATION

This section is discussing the simulation environment, experiments nature, dataset, performance parametric and detail results. The experiment in this paper is simulated and verified with the MATLAB tool 2018b, it can handle one-dimensional, two-dimensional and multi-dimensional arrays well and display them graphically [33-36]. This paper uses the powerful graphics processing function of MATLAB to verify the superiority of LEACH-CR protocol in energy consumption. In this research paper, we discuss the LEACH-DSEE, LEACH & LEACH-C & LEACH-M are simulated and validated by using MATLAB 2081b tools, and the performance of the four protocols is analyzed and compared by means of the graphical display. The 200 sensor nodes are erratically distributed in the monitoring area of two hundred by two hundred meters and the BS is far away from the monitoring area. The experiment time duration or number of rounds 1600 sec.

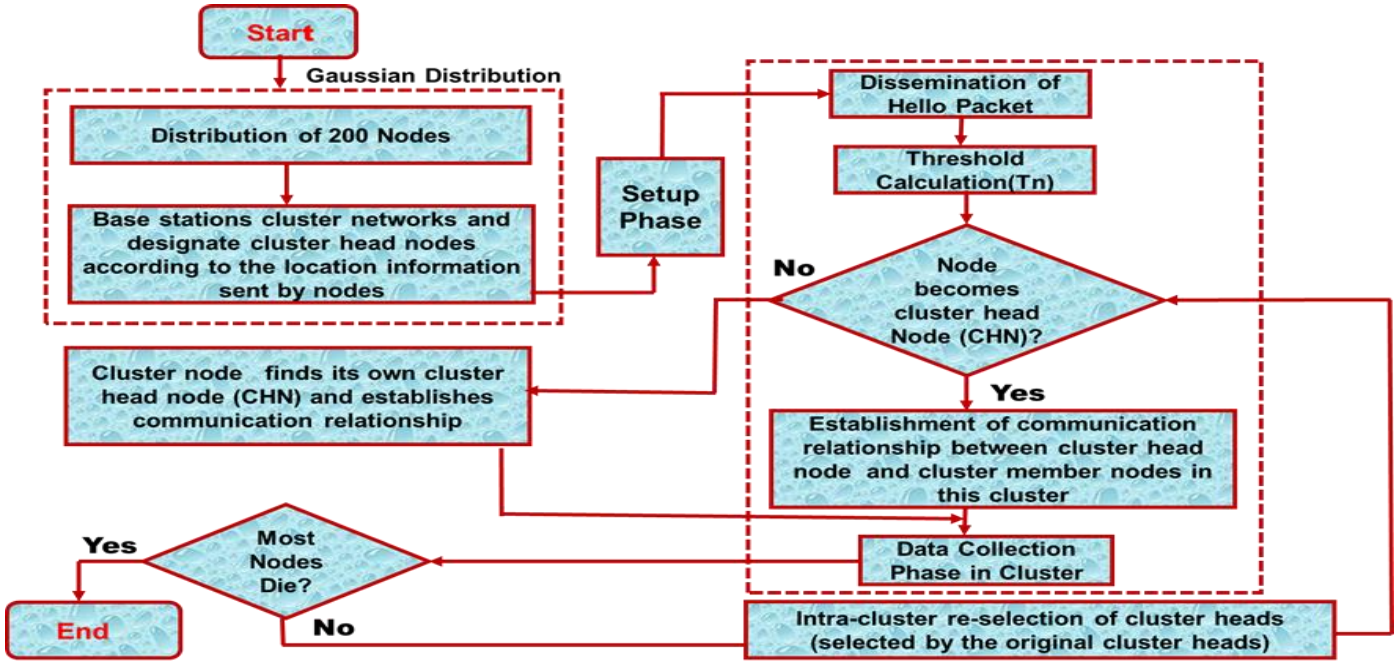

Start
Gaussian Distribution
Distribution of 200 Nodes
Base stations cluster networks and designate cluster head nodes according to the location information sent by nodes
Setup Phase
Dissemination of Hello Packet
Threshold Calculation(Tn)
No
Node becomes cluster head Node (CHN)?
Yes
Cluster node finds its own cluster head node (CHN) and establishes communication relationship
Establishment of communication relationship between cluster head node and cluster member nodes in this cluster
Data Collection Phase in Cluster
Yes
Most Nodes Die?
No
End
Intra-cluster re-selection of cluster heads (selected by the original cluster heads)

Fig. 5. Algorithm for the Selection of Optimal CHN.

A. Function Definition

In the simulation experiments of LEACH-DSEE, LEACH and LEACH-C and LEACH-M, many functions are defined. They work together to establish the network topology and data collection using LEACH-DSEE, LEACH and LEACH-C & LEACH-M protocols. The important functions in the simulation code are as follows:

1) Init (): Initialize all sensor nodes of the protocol, including location information, initial energy, and common member nodes.

2) Build cluster (): Network clustering. The clustering processes of LEACH, LEACH-DSEE, and LEACH-C & LEACH-M protocols are different from each other.

3) Gather data (): Data collection. The node collects data and transmits it to BS. Cluster node transmits the collected data directly to the CHN, which fuses the collected data and sends it to the B.S.

B. Simulation Environment

In equation 3, d is the distance difference b/w source and BS. When d is smaller than threshold d_0, the communication model of the node is free space channel.

$$E_{tx}(d,k) = \begin{cases} K\times E_{elec}+k\times \varepsilon_{fs}\times d^2 & d<d_0 \\ K\times E_{elec}+k\times \varepsilon_{amp}\times d^4 & d\geq d_0 \end{cases} \quad (3)$$

When d is not less than the threshold d_0, the multi-channel attenuation model is adopted. The consumption of energy in s-nodes for receiving data (kbit) in Equation (4).

$$E_{rx}(k) = k \times E_{elec} \quad (4)$$

In equation (3) and (4), E_{elec} is the consumption loss parameter of the circuit, and the coefficients of the free space (ε_{fs}) channel and ε_{amp} multiple channel attenuation model are and, respectively. The parameters of the experimental environment are as follows (Table II):

TABLE II. ENVIRONMENTAL PARAMETERS OF A SIMULATION EXPERIMENT

Parameter term	**Parameter values**
Total number of nodes	200 Individual
Monitoring regional scope	250m x 250m
Base station location	(250,500)
Packet size	4000 bits
Controlling package size	100 bits
The initial energy of nodes	0.5J
Energy usage of DF (nJ/bits)/message	5
E_{elec}	50
ε_{fs}	10
ε_{amp}	0.013
d_0	87m
LEACH-DSEE Protocol Partition Supervision Block Number of Measured Areas	9 Block
LEACH-DSEE protocol Threshold M &Rounds	10 individuals 1600

C. Performance Assessment

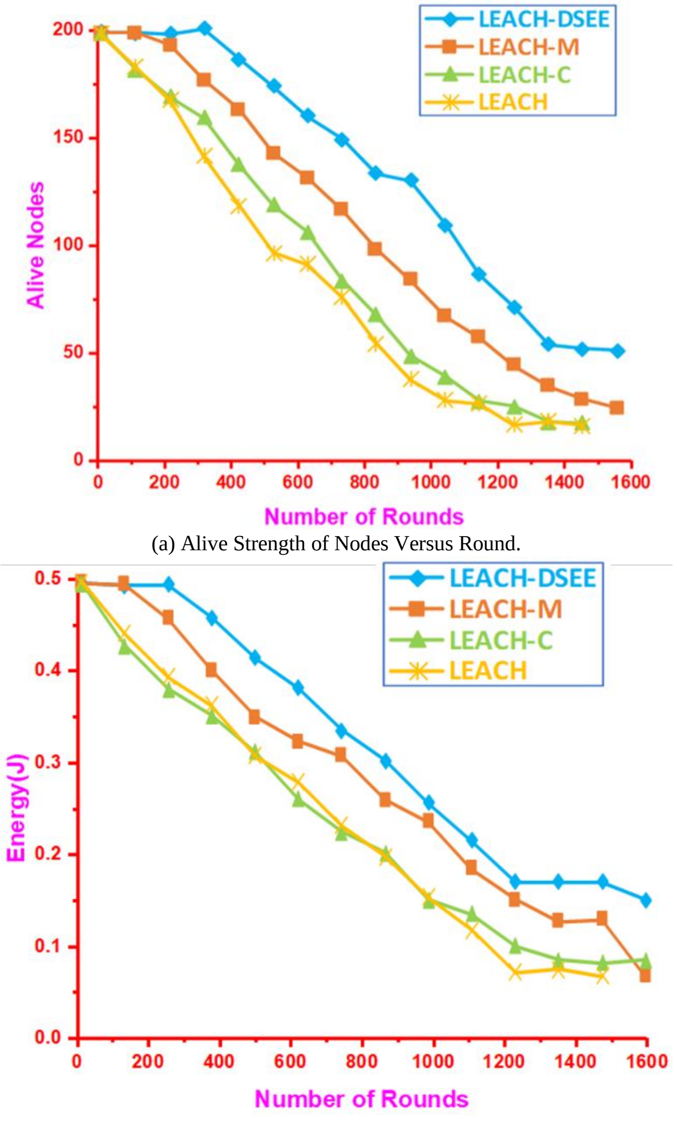
The performance assessment of an RP is based on the number of dead rounds of 1st the whole network. If the later the node starts to die, the more concentrated the time of node death, the better the performance of this protocol. If the amount of residual energy of a node is zero, then the node is measured as dead. If the strength of the dead node in the network is more than 85% of the total strength of nodes, the network considered as dead. Therefore, this study uses the number of dead rounds of the first node and 85% of the nodes in the whole network to judge the performance of a routing protocol [34, 40].

The simulation results of four RP LEACH-DSEE, LEACH and LEACH-C and M show that LEACH-DSEERP is superior to the other four in terms of network lifetime and residual energy. Fig. 6 is representing the simulation of the changes of alive nodes in LEACH-DSEE RP, LEACH RP & LEACH-C and LEACH -M with the increase of running rounds Energy comparison. Fig. 6(a) shows that the Maximum alive node in various LEACH after the 1600 rounds. It can be concluded that LEACH-DSEE RP is superior to LEACH &LEACH-C and LEACH-M RP because deaths rate is very low. Fig. 6(b) is a comparison of the total network energy remain in LEACH-DSEE RP as compare to the LEACH and LEACH-C and LEACH-M protocols. As can be seen from Fig. 6(b), the total remained energy in LEACH-DSEE RP network nodes is much higher than as compare to LEACH and LEACH-C and LEACH-M protocols with the increase of running rounds. Therefore, LEACH-DSEE RP is superior to LEACH and LEACH-C and LEACH-M RP in terms of network lifetime and residual energy. This is of CHN more uniform through the method of regional clustering Secondly, in LEACH-DSEE, each cluster chooses the node with the largest residual energy as CHN, which balance the network energy consumption and avoids the premature death of the node with little residual energy because it acts as the CHN. Finally, the CHN away from the BS sends data to the BS by multi-hop method which solves the problem of excessive energy consumption. The proposed LEACH-DSEE introduces connectivity facility among the sensing nodes, therefore CHN requires low consumption of energy to communicate with family nodes, which results in maximum strength of remained alive nodes. Fig. 6(b) showing the remained energy versus rounds and 200 nodes. This result shows the maximum remain energy in LEACH-DSEE as compare to another leach RP.

Fig. 7(a) shows the ratio of packet delivery between the strength of node and during the packet delivery height ratio provide our algorithm LEACH-DSEE RP. Fig. 7(b) shows the energy spectrum during the transmission.

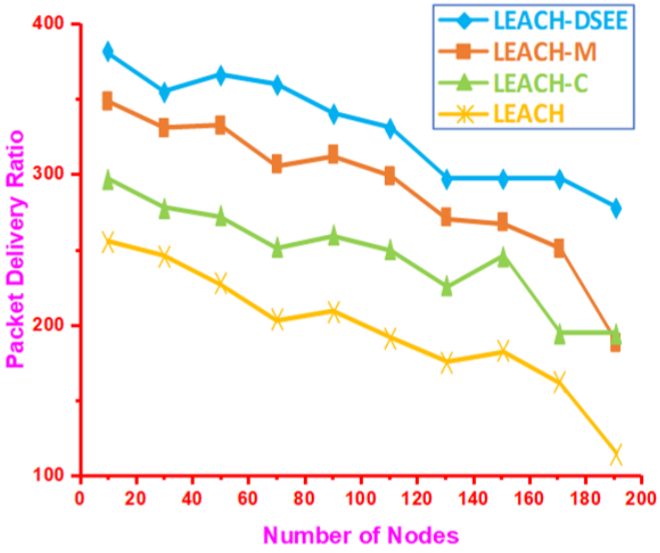
D. Comparison of LEACH-DSEE RP with LEACH, LEACH-M &LEACH-C RP

Table III is representing a comparison between LEACH-DSEE routing protocol with the other four LEACH protocol. LEACH-DSEE RP mainly improves around CHN. When choosing CH, residual energy, the distance between CH and the BS, distribution, and number of CH are considered. However, neither protocol can be applied to large-scale WSNs networks. The study the LEACH Protocol Improvement for very large-scale WSNs network applications.



(a) Alive Strength of Nodes Versus Round.

Fig. 6. (b) Energy Comparison.



(a) Ratio of Packet Delivery b/w No. of Nodes.

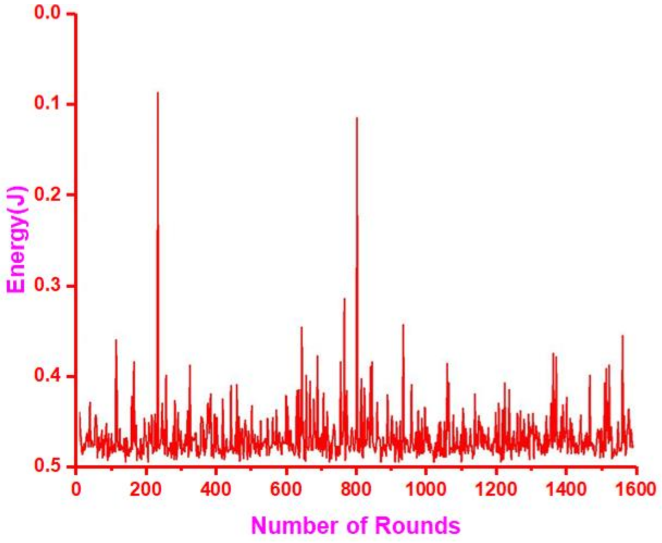


Fig. 7. (b) Energy use Spectrum during the Transmission.

TABLE III. COMPARISON OF LEACH-DSEE RP WITH LEACH, LEACH-M &LEACH-C RP

	Whether to Consider Clusters Remaining of HN Residual energy	**Whether to Solve CHN Distance from BS Problem**	**Whether to Consider CH Distribution of nodes and Number**	**Can it be applied? On a Very-large-scale Network Luo Li**
LEACH, LEACH-M LEACH-C RP Agreement	No	No	No	No
LEACH-DSEE Agreement	Yes	Yes	Yes	Yes

VII. CONCLUSION

As a classical hierarchical routing protocol, LEACH solves the shortcomings of a planar routing protocol such as high energy consumption and long delay. But the LEACH protocol not suitable for very large-scale WSNs. The LEACH- DSEE protocol establishes several data communication links to the BS through the CHN layer, so the LEACH- DSEE protocol can be used in very large-scale wireless sensor networks. Experiments results of MATLAB 2018b tools show that LEACH-DSEE protocol can greatly improve the life cycle of wireless sensor networks and reduce the energy consumption of sensor networks. In the proposed scheme, the selection of CHN is based remained energy & probabilistic connectivity between family nodes. The improved choice of CHN increases remaining energy of sensing nodes and increase network lifetime.

ACKNOWLEDGMENT

This Research work is done by the Chinese Government scholarship student of Master of Electronics and communication.

References

[1] Rahman, K. C. (2010). A survey on sensor network. Journal of Computer and Information Technology, 1(1), 76-87.

[2] Limin, S., Jianzhong, L., Yu, C., & Hongsong, Z. (2005). Wireless sensor networks. Beijing: Tsinghua University Press, 5, 7-8.

[3] Cui, L., Ju, H., & Miao, Y. (2005). Research progress of wireless sensor network. Journal of Computer Research and Development., 42(l), 163-174.

[4] Liu, J., Bic, L., Gong, H., & Zhan, S. (2016). Data collection for mobile crowdsensing in the presence of selfishness. Eurasip Journal on Wireless Communications & Networking, 2016(1), 82.

[5] Heinzelman, W. R., Chandrakasan, A., & Balakrishnan, H. (2000, January). Energy-efficient communication protocol for wireless microsensor networks. In Proceedings of the 33rd annual Hawaii international conference on system sciences (pp. 10-pp). IEEE.

[6] Yang, H., Zhou, L., He, K., Deng, C., Zhao, X., & Qiu, Z. (2010, July). A probabilistic QoS model-checking for the dynamic routing protocol. In 2010 10th International Conference on Quality Software (pp. 441-448). IEEE.

[7] Nacer, A., Marhic, B., & Delahoche, L. (2017, May). Smart Home, Smart HEMS, Smart heating: An overview of the latest products and trends. In 2017 6th International Conference on Systems and Control (ICSC) (pp. 90-95). IEEE.

[8] Kabilan, K., Bhalaji, N., Selvaraj, C., Mahesh, K. B., & Karthikeyan, P. T. R.. Performance analysis of IoT protocol under different mobility models. Computers & Electrical Engineering.

[9] Li, L. Y., & Liu, C. D. (2015). An Improved Algorithm of LEACH Routing Protocol in Wireless Sensor Networks. International Conference on Future Generation Communication & Networking.

[10] Rani, R., Kakkar, D., Kakkar, P., & Raman, A. (2019). Distance-Based Enhanced Threshold Sensitive Stable Election Routing Protocol for Heterogeneous Wireless Sensor Network.

[11] Zhang, W. H., La-Yuan, L. I., Zhang, L. M., & Wang, X. Z. (2008). Energy consumption balance improvement of leach of wsn. Chinese Journal of Sensors & Actuators, 21(11), 1918-1922.

[12] Attea, BA, & Khalil, EA (2012). A new evolutionary-based routing protocol for clustered heterogeneous wireless sensor networks. Applied Soft Computing Journal, 12 (7), 1950-1957.

[13] Ma, K., Zhang, Y. , & Trappe, W.. (2008). Managing the mobility of a mobile sensor network using network dynamics. IEEE Transactions on Parallel and Distributed Systems, 19(1), 106-120.

[14] Yang Caixia, Koryo, Luchang. A Clustering Algorithm for Balancing Node Energy Consumption in WSN [J]. Journal of Jiamusi University (Natural Science Edition), 2015, 33 (6): 925-928.

[15] Du Chao. Analysis and Simulation of LEACH-C protocol based on NS2 [J]. Electronic measurement technology, 2011, 34 (9): 121-123.

[16] Kwiatkowska, M., Norman, G., Parker, D., & Qu, H. (2010). Assume-guarantee verification for probabilistic systems. Lecture Notes in Computer Science, 6015, 23-37.

[17] Singh, C. P., Vyas, O. P., & Tiwari, M. K. (2008). A Survey of Simulation in Sensor Networks. International Conference on Computational Intelligence for Modelling Control & Automation.

[18] J. Zhu, R. Pecen. A novel automatic utility data collection system using IEEE 802.15.4-compliant wireless mesh network[C]. In proceedings of IAJC-IJME International Conference, 2008.

[19] [Aslam, M., Javaid, N., Rahim, A., Nazir, U., Bibi, A., & Khan, ZA (2012). Survey of extended leach-based clustering routing protocols for wireless sensor networks.

[20] Murthy, S., & Garcialunaaceves, JJ (1996). An efficient routing protocol for wireless networks. Mobile Networks & Applications, 1 (2), 183-197.

[21] Rani, R., Kakkar, D., Kakkar, P., & Raman, A. (2019). Distance-Based Enhanced Threshold Sensitive Stable Election Routing Protocol for Heterogeneous Wireless Sensor Network.

[22] Zheng, X., Ge, L., Guo, W., & Liu, R. (2007). Cross-Layer Design and ant-colony optimization-based load-balancing routing protocol for ad-hoc networks. Frontiers of Electrical and Electronic Engineering in China, 2(2), 219-229.

[23] Ortolani, S., Conti, M., Crispo, B., & Pietro, R. D. (2011). Events privacy in WSNs: A new model and its application. IEEE International Symposium on A World of Wireless.

[24] Wang, Q. H., Guo, H. Y., & Ji, Y. H. (2011). Research and improvement of leach protocol for wireless sensor networks. Journal of Chongqing University of Posts & Telecommunications, 562-564, 1304-1308.

[25] Yanchun, T. U., & Guo, A. (2006). Routing algorithms and simulation of the wireless sensor network. Computer Engineering, 32(22), 124-126.

[26] Feng, Y., Kang, H. E., Yang, H., Qiu, Z., & Liu, Y. (2014). The research and optimization of data gathering protocol for wireless sensor network. Chinese Journal of Sensors & Actuators, 27(3), 355-360.

[27] Hüseyin Özgür Tan. (2003). Power-efficient data gathering and aggregation in wireless sensor networks. Acm Sigmod Record, 32(4), 66-71.

[28] Wu, X., & Sheng, W. (2010). Performance Comparison of LEACH and LEACH-C Protocols by NS2. International Symposium on Distributed Computing & Applications to Business.

[29] Thein, M. C. M., & Thein, T. (2010, January). An energy-efficient cluster-head selection for wireless sensor networks. In 2010 International Conference on Intelligent Systems, Modelling, and Simulation (pp. 287-291). IEEE.

[30] Li Lanying, Liu Changdong. An improved LEACH algorithm for wireless sensor network routing protocol [J]. Journal of Harbin University of Technology, 2015, 20 (02): 75-79.

[31] Mankita, Singh, E. P., & Rani, E. S. Improved leach routing communication protocol for wireless sensor network. International Journal of Distributed Sensor Networks, 2012, 1-6.

[32] Afsar, M. M., & Tayaranin, M. (2014). Clustering in sensor networks: a literature survey. Journal of Network & Computer Applications, 46, 198-226.

[33] Wang, Y. H., Yu, C. Y., Chen, W. T., & Wang, C. X. (2008). An average energy-based routing protocol for Mobile Sink in wireless sensor networks. IEEE International Conference on Ubi-media Computing.

[34] Qin, Xianglin, H. Zhang, and Y. Zhang. "Research on wireless sensor networks clustering routing algorithm based on energy balance." International Conference on Measurement 2012.

[35] Sharma, N., Prakash Verma, C., & Kumar Mahirania, R. (2014). Prototypedleach routing algorithm for the enhancement of energy efficiency in the wireless sensor network. International Journal of Computer Applications, 95(18), 30-32.

[36] Aslam, M., Javaid, N., Rahim, A., Nazir, U., Bibi, A., & Khan, ZA (2012). Survey of Extended LEACH-Based Clustering Routing Protocols for Wireless Sensor Networks. IEEE International Conference On High-Performance Computing & Communication & IEEE International Conference on Embedded Software & Systems.

[37] Bartolini, N., Massini, A., & Silvestri, S. (2012). P&p: an asynchronous and distributed protocol for mobile sensor deployment. Wireless Networks, 18(4), 381-399.

[38] Arya, R., & Sharma, S. C. (2018). Energy optimization of, energy-aware routing protocol and bandwidth assessment for wireless sensor network. International Journal of System Assurance Engineering & Management, 26(5), 1-8.

[39] Di, T., Li, T., Jian, R., & Jie, W. (2015). Cost-aware secure routing (caser) protocol design for wireless sensor networks. Parallel & Distributed Systems IEEE Transactions on, 26(4), 960-973.

[40] Al-Karaki, J. N., & Gawanmeh, A. (2017). The optimal deployment, coverage, and connectivity problems in wireless sensor networks: revisited. IEEE Access, PP(99), 1-1.

[41] Babar Ali,Tariq Mahmood, Muhammad Abbas, Muzamil Hussain, Habeeb Ullah, Anupam Sarker, Asad Khan.(2019).LEACH Robust Routing Approach Applying Machine Learning.IJCSNS,.19(6),18-26.

Analysis of Software Tools for Longitudinal Studies in Psychology

Pavel Kolyasnikov[1]
Russian Academy of Education
Moscow, Russia

Evgeny Nikulchev[2], Vladimir Belov[3], Anastasia Silaeva[4]
MIREA–Russian Technological University, Moscow, Russia

Alexander Kosenkov[5]
Sechenov First Moscow State Medical University, Moscow, Russia

Artem Malykh[6]
National Research Nuclear University MEPhI, Moscow, Russia

Zalina Takhirova[7]
Saint-Petersburg State University
Saint-Petersburg, Russia

Sergey Malykh[8]
Lomonosov Moscow State University, Moscow, Russia

Abstract—**Longitudinal studies allow to access the review of causal hypotheses directly. It means that they make possible causal relation between the order of impacts (i.e. life events, educational effects, etc.) and the consequences that then occur. Long-term data storage has specific requirements for software and methods of data storage and conversion. The paper introduces criteria for evaluating software tools in the context of their application in longitudinal studies in psychology. The study is devoted to the analysis of popular tools of psychological research based on the criteria, which were introduced.**

Keywords—Software tools for psychological research; choice of the tools for psychological research; longitudinal studies; criteria for software evaluation

I. Introduction

A longitudinal study is a type of survey aimed to study the group of people (research sample) that is monitored for a certain period of time. Longitudinal surveys usually include the repeated measurements of different variables (from health status to value orientations). For example, a study of the school-age children development may include annual measurements of intelligence, math skills and reading skills to study cognitive development and learning abilities or impaired learning abilities.

Longitudinal studies are difficult to conduct, but they provide a number of advantages in comparing with other research methods:

1) Assessment of individual changes;

2) Identification of environmental events initiating certain processes;

3) A prospective identification of environmental impacts (i.e., an assessment of the order of impacts and consequences over time);

4) Separation of time effects: cohort, period, age;

5) Cohort effects control.

Longitudinal data are key in educational sciences, because they allow to evaluate the cause-effect relationships between different educational influences and their influence on the development of a child.

It is no coincidence that large-scale longitudinal studies all over the world occupy an important place not only in fundamental science, but also in educational practice. For example, the famous longitudinal study of the effectiveness of the preschool education program on the further personal achievements of its students (HighScope program [1]) served as a basis for policy makers in many countries to make crucial management decisions, in particular, to transform scattered preschool services into preschool education. One of the oldest longitudinal studies of 1000 intellectually gifted children [2] was started in 1921 and is still ongoing. The longitudinal study of 13,687 humans born from March 3 to March 9, 1946 in England, Wales and Scotland has been going on for over 70 years. The results of this study gave a number of important results for understanding not only the physical, but also the cognitive development of children. Thus, 80% of this sample is over 70 years old and the data of these people currently help to determine the factors contributing to the preservation of physical and mental health, cognitive skills, social and psychological well-being in old age.

Interdisciplinary research allows to identify a full range of different factors not only of mental development disorders, but also of personality formation, motivation, value orientations and success in learning educational programs. These factors may be associated with the characteristics of prenatal and early child development, morphofunctional maturation of the brain and the systemic brain organization of cognitive functions, language environment, genetic profile, lifestyle, etc.

The results of longitudinal interdisciplinary research provide huge data sets, that opens for new opportunities to use the intelligent analysis methods for assessing both global and local systemic effects (for example, assessing the impact of various educational technologies on the mental development of children), as well as to understand the characteristics of individual response on these factors. Without big data analyzing, it is impossible to research fundamental childhood, scientifically based administrative and managerial decisions.

The use of Web technologies significantly expands opportunities for psychological research. For such usage organization, it is important to solve the problem of the tools for psychological research choice. The paper is devoted to the analysis of popular tools for psychological research. The second part introduces evaluation criteria and analyzing of software. The third one describes the software development based on the analysis of functionality.

II. Criteria for Software Evaluation

To analize existing software solutions used for psychological researches, it is required to structure the information in such a way that it allows to compare different tools.

The list of criteria used to evaluate software for psychological researches is described below:

- **Web-site**–web-site address of software solution in Internet;
- **License**–term of use, cost and type of license (commercial, free license, open source etc.);
- **Platform**–list of supported platforms on which the application runs (web browser, web-service SaaS (Software as a Service), desktop, mobile application etc.);
- **Country of production–**country in which the application was developed;
- **Functional purpose–**what tasks the application is focused on, how developers are positioning their tool;
- **Test types**–the types of tests supported by application (questionnaires, cognitive tests);
- **Data storage resource**–the resource storing data (local data storage, remote data storage, SaaS storage (Software as a Service));
- **Country of data storage**–location of application servers for data storage (for SaaS solutions);
- **Personal data storage features**–techniques of data collection, techniques of users identification, the need of researches registration, opened and closed access to researches completion;
- **Techniques of tests creation–**which tests creation techniques are supported by application (tests constructors, tests description format, programming);
- **Test playing**–support opportunities of tests reuse, coping and portability;
- **Programming knowledge–**the need of programming skills for application use and test creation;
- **Complexity**–complexity of using the application, what is needed for start use, the problems that may arise, interface features;
- **Tools for data analyses**–existing tools for data processing and research results analyses;
- **Opportunities for data export–**techniques and formats of test results storing (.txt, .csv etc.);
- **Integration tools**–opportunities for functionality expansion supported by application (API, modules, plug-ins, integration with third-party solutions etc.);
- **Solutions relevance**–application updates, compliance with current trends and standards;
- **Solutions features–**features of application and what to take note of.

Functionality opportunities described systems was analyzed and checked by authors.

A. PsychoPy

PsychoPy is an open source project that allows to carry out experiments in the field of neurobiology, psychology and psychophysics [3, 4]. PsychoPy supports many constructors of psychological tests: Builder (visual test constructor shown in the Fig. 1) and Coder (test programming).

By start the experiment from visual constructor, PsychoPy translates the code into Python code and then executes is. The code written in PsychoPy Builder can be saved separately and then executed. It should be noticed that backward compatibility does not exist. It means that there is no opportunity to create visual view from the code. In the presence of programming skills it is possible to develop any types of tests.

PsychoPy supports its own format of experiment file (.psyexp) that is based on xml format and displays the structure of experiments built using the Builder GUI, which generally makes it platform independent.

PsychoJS [5] was created in 2016 as a project that provides an opportunity to run researches developed in PsychoPy Builder in web browser. It should be noticed that PsychoPy cannot export all possible experiments into PsychoJS scripts. Researches created using standard components will work, but the researches created using third-party components or code won’t. But according to completed tests and feedback, PsychoJS does not work well and has a number of problems.

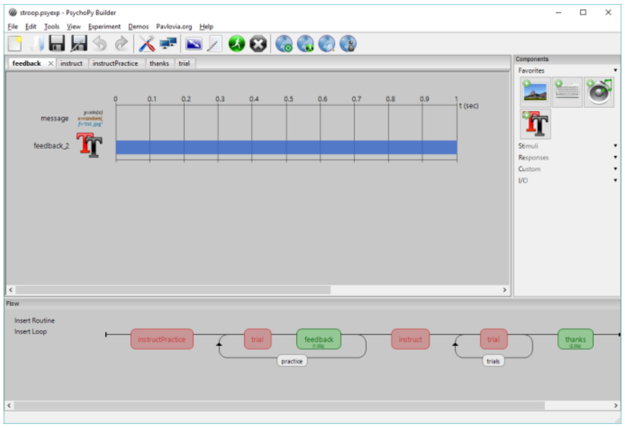


Fig. 1. Visual Test Constructor PsychoPy Builder.

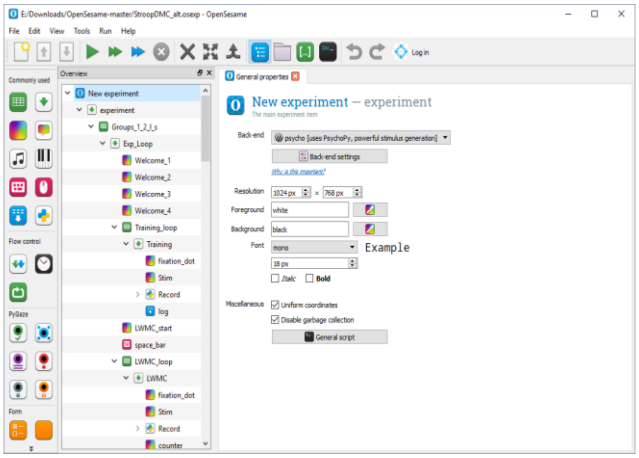
The detailed characteristics of PsychoPy are presented below:

- **Web-site**–http://www.psychopy.org/
- **License**–open source (GNU GPL 3+)
- **Platform**–desktop (Windows, MacOS, Linux), web browser (based on PsychoJS)
- **Country of production**–Great Britten (with the support of the University of Nottingham)
- **Functional purpose**–experiments in the field of neurobiology and psychology
- **Test types**–cognitive tests
- **Data storage resource**–own local storage
- **Country of data storage**–no
- **Personal data storage features**–data is collected and stored locally, researcher is responsible for its storing
- **Techniques of tests creation**–test constructor, description format, programming
- **Test playing**–yes (format of .psyexp)
- **Programming knowledge**–not required for basic use of test constructor
- **Complexity**–installation is carried out using the installer or installation package, the interface is easy, some actions require reading documentation
- **Tools for data analyses**–no
- **Opportunities for data export**–data export in .csv format
- **Integration tools**–API support, integration with outside research tools
- **Solutions relevance**–popular and relevant, beeing supported and developed
- **Solutions features**–visual editor generates code in Python, support of outside tools

B. *OpenSesame*

OpenSesame is a tool for experiments creation and running in the field of psychology, neurobiology and experimental economics [6, 7]. The application contains an advanced and multifunctional graphical test editor, which is shown in Fig. 2.

The main feature of OpenSesame is the Backend, which is a software level that processes the input (keyboard, mouse, etc.) and output (presentation presentation, sound playback, etc.). There is many libraries that provide this type of functionality, OpenSesame can use every type. Currently there is four type of Backend: Legacy (PyGame functionality is a set of libraries for developing computer games and multimedia applications), Psycho (PsychoPy functionality), Droid (Android functionality) and Xpyriment (Expyriment functionality).



Back-end: psycho [uses PsychoPy, powerful stimulus generation]

Resolution: 1024 px × 768 px

Foreground: white

Background: black

Font: mono, 18 px

□ Italic □ Bold

Miscellaneous: ☑ Uniform coordinates ☑ Disable garbage collection

Fig. 2. The Main Test Constructor Window of OpenSesame.

Characteristics of OpenSesame:

- **Web-site**–http://osdoc.cogsci.nl/
- **License**–open source (GNU GPL 3)
- **Platform**–desktop (Windows, MacOS, Linux), modile (Android)
- **Country of production**–Holland
- **Functional purpose**–experiments in the field of neurobiology and psychology
- **Test types**–cognitive tests
- **Data storage resource**–local storage
- **Country of data storage**–no
- **Personal data storage features**–data is collected and stored locally, researcher is responsible for its storing
- **Techniques of tests creation**–test constructor, description format, programming
- **Test playing**–yes (.osexp format)
- **Programming knowledge**–not required for basic use of test constructor
- **Complexity**–installation is carried out using the installer or installation package, the interface is easy, some actions require reading documentation
- **Tools for data analyses**–no
- **Opportunities for data export**–data export in format of .csv
- **Integration tools**–API, OpenSesame Script (simple language for experiment description), plug-ins and extensions (package manager pip), integration with outside research tools (Mouse tracking, Emotiv EEG, Oculus rift etc.), integration with Open Science Framework (OSF)
- **Solutions relevance**–popular and relevant, beeing supported and developed

- **Solutions features**–application is a superstucture (backend) over other tools (PyGame, PsychoPy и Expyriment), quality documentation.

C. jsPsych

jsPsych is not complete application, but is library on JavaScript for web browser. It is developed for behavior experiments creation [8]. Thus, programming knowledge is required for experiment creation using JavaScript libraries. Fig. 3 shows main page of jsPsych.

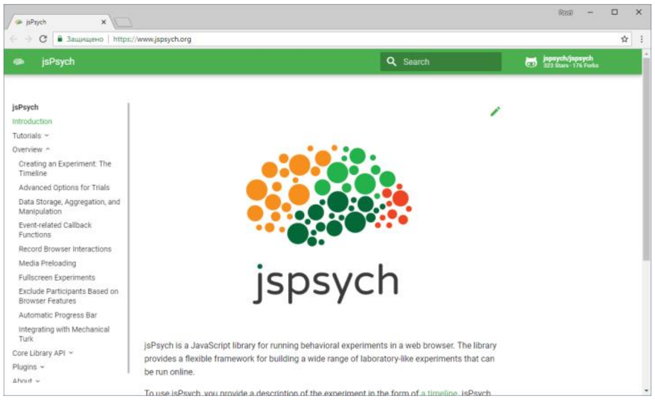
jsPsych provides a framework for experiments development using a set of flexible plug-ins that create different tasks. Using different plug-ins combination, it is possible to create many types of experiments. It should be noticed that jsPsych library is a part of some software solutions or can be used as extension and is recommended to use.

Characteristics of jsPsych:

- **Web-site–**http://www.jspsych.org/
- **License–**open source (MIT)
- **Platform–**web browser
- **Country of production–**USA
- **Functional purpose–**JavaScript library for cognitive tests creation
- **Test types–**questionnaires, cognitive tests (on the basis of JavaScript it is possible to create every test)
- **Data storage resource–**depending on the code
- **Country of data storage–**no
- **Personal data storage features–**data is stored depending on the code
- **Techniques of tests creation–**programming
- **Test playing–**no
- **Programming knowledge–**required for test creation
- **Complexity–**it is required to learn documentation of library and write code on JavaScript
- **Tools for data analyses–**no
- **Opportunities for data export–**no
- **Integration tools–**use of third-party libraries on JavaScript
- **Solutions relevance–**being supported and updated
- **Solutions features–**tools is not application, but library, the functionality should be developed by researcher.

D. PEBL

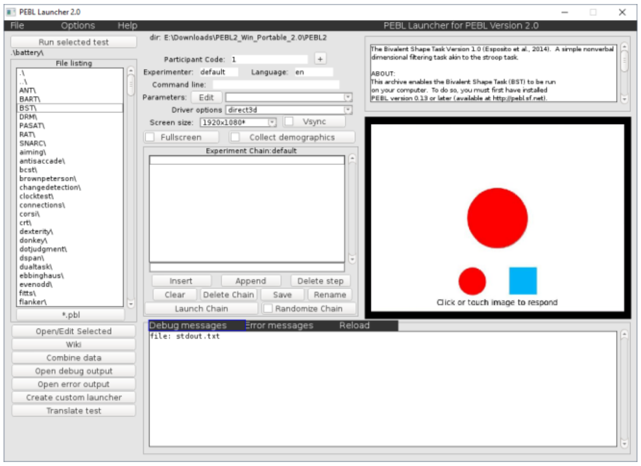
PEBL (Psychology Experiment Building Language) is free software that realize psychological test batteries [9, 10]. They are designed to develop a wide range of computer psychological tests that are of interest for neuropsychometric, cognitive and clinical community [11]. Fig. 4 shows main window of PEBL.


jsPsych
Защищено | https://www.jspsych.org
jsPsych
Search
jspsych/jspsych
323 Stars · 176 Forks
jsPsych
Introduction
Tutorials
Overview
Creating an Experiment: The Timeline
Advanced Options for Trials
Data Storage, Aggregation, and Manipulation
Event-related Callback Functions
Record Browser Interactions
Media Preloading
Fullscreen Experiments
Exclude Participants Based on Browser Features
Automatic Progress Bar
Integrating with Mechanical Turk
Core Library API
Plugins
About
jspsych

jsPsych is a JavaScript library for running behavioral experiments in a web browser. The library provides a flexible framework for building a wide range of laboratory-like experiments that can be run online.

To use jsPsych, you provide a description of the experiment in the form of a timeline. jsPsych

Fig. 3. Main page of jsPsych.



dir: E:\Downloads\PEBL2_Win_Portable_2.0\PEBL2

.\battery\

```
.\
..\
ANT\
BART\
BST\
DRM\
PASAT\
RAT\
SNARC\
aiming\
antisaccade\
bcst\
brownpeterson\
changedetection\
clocktest\
connections\
corsi\
crt\
dexterity\
donkey\
dotjudgment\
dspan\
dualtask\
ebbinghaus\
evenodd\
fitts\
flanker\
```

Participant Code: 1

Experimenter: default   Language: en

Command line:

Parameters:

Driver options: direct3d

Screen size: 1920x1080*

Experiment Chain:default

The Bivalent Shape Task Version 1.0 (Esposito et al., 2014). A simple nonverbal dimensional filtering task akin to the stroop task.

ABOUT:
This archive enables the Bivalent Shape Task (BST) to be run on your computer. To do so, you must first have installed PEBL version 0.13 or later (available at http://pebl.sf.net).

```
file: stdout.txt
```

Fig. 4. Main Window of PEBL.

PEBL offers a simple programming language for creating and conducting many standard experiments written in C++. To launch the test, it is necessary to choose it in the catalog, set up and run. Hidden launch is not provided. PEBL launcher is used for test running (Fig. 4).

It should be noted that PEBL has some problems with support of Cyrillic. Thus, ASCII characters are use default. To resolve the issue following is required: use the font available in PEBL with Cyrillic support; repeatedly save PEBL script in UTF-8 format without BOM (UTF-16 is not supported) using corresponding text editor.

Although the application is being still supported, it is recommended to use modern solutions (i.e., PsychoPy or OpenSesame described earlier in this paper).

Characteristics of PEBL:

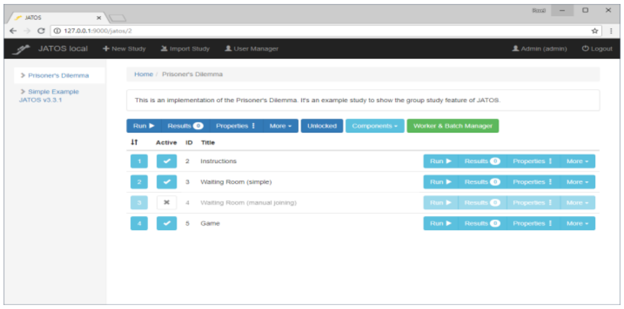
- **Web-site–**http://pebl.sourceforge.net/
- **License–**open source (GNU GPL 2)
- **Platform–**desktop (Windows, MacOS, Linux)
- **Country of production–**USA

- **Functional purpose–**experiments in the field of psychology
- **Test types–**cognitive tests
- **Data storage resource–**local storage
- **Country of data storage–**no
- **Personal data storage features–**data is collected and stored locally, researcher is responsible for its storing
- **Techniques of tests creation–**programming
- **Test playing–**yes (test is created in the own .pbl format)
- **Programming knowledge–**required for test development
- **Complexity–**installation is carried out using the installer or installation package, interface is not convenient and does not meet the modern requirements
- **Tools for data analyses–**no
- **Opportunities for data export–**data export in .txt and .csv format
- **Integration tools–**no
- **Solutions relevance–**interface looks outdated, application is being supported and used
- **Solutions features–**there is a library with prepared tests, there are problems with support of Cyrillic.

E. JATOS

JATOS (Just Another Tool for Online Studies) is developed for creation and completion of online researches on the own server (local or remote) [12]. It means the researcher saves full control over the access to test results. Fig. 5 shows interface of JATOS.

JATOS is client-server application based on server and the work with application is carried out in the browser. There is 2 use-cases: start the server locally and install it on the server. An installation package for Windows, MacOS and Linux is available for local installation. For installation on server, JATOS has another package that does not contain JAVA and a configured server for a concrete OS. The actual JATOS instance on the server is not different from the local one, the main difference is a need to configure manually the server and install necessary libraries.



This is an implementation of the Prisoner's Dilemma. It's an example study to show the group study feature of JATOS.

↓↑	Active	ID	Title
1	✓	2	Instructions
2	✓	3	Waiting Room (simple)
3	✗	4	Waiting Room (manual joining)
4	✓	5	Game

Fig. 5. Interface of JATOS.

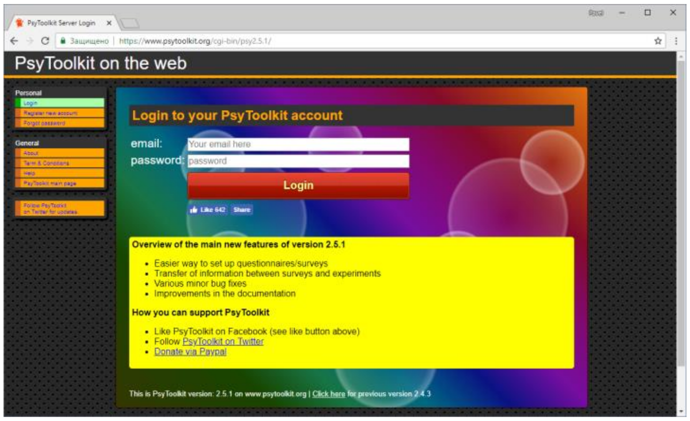
Tests is being developed manually with the help of HTML, CSS and JavaScript. The use of jsPsych is proposed for developers.

Characteristics of JATOS:

- **Web-site–**https://www.jatos.org/
- **License–**open source (Apache 2 License)
- **Platform–**web browser, server (Windows, MacOS, Linux)
- **Country of production–**no information
- **Functional purpose–**a tool for psychological online researches
- **Test types–**questionnaires, cognitive tests (it is possible to create any tests based on JavaScript)
- **Data storage resource–**local storage
- **Country of data storage–**no
- **Personal data storage features–**data is collected and stored locally, researcher is responsible for its storing
- **Techniques of tests creation–**programming (JavaScript code along with setting in the application interface)
- **Test playing–**yes (export and import from zip format)
- **Programming knowledge–**required for test creation and updating
- **Complexity–**server is required to run, it is necessary to develop every page of test on JavaScript
- **Tools for data analyses–**no
- **Opportunities for data export–**data export in txt format
- **Integration tools–**use of third-party libraries on JavaScript, integration with Amazon Mechanical Turk (MTurk), Prolific and analogues
- **Solutions relevance–**application is being supported and updated
- **Solutions features–**group researches supported, there is library of prepared research examples, use of the jsPsych library, support of researches with a number of researchers (i.e., Prisoner's dilemma).

F. PsyToolkit

PsyToolkit is free tool for creation and completion of questionnaires and cognitive-psychological experiments and tests [13, 14]. PsyToolkit is used by researches and students all over the worlds. They use it for their own projects. PsyToolkit has many versions: online (in web browser) and offline (Linux version). Fig. 6 shows main page of PsyToolkit web version.



Login to your PsyToolkit account

Overview of the main new features of version 2.5.1

- Easier way to set up questionnaires/surveys
- Transfer of information between surveys and experiments
- Various minor bug fixes
- Improvements in the documentation

How you can support PsyToolkit

- Like PsyToolkit on Facebook (see like button above)
- Follow PsyToolkit on Twitter
- Donate via Paypal

This is PsyToolkit version: 2.5.1 on www.psytoolkit.org | Click here for previous version 2.4.3

Fig. 6. Main Page of PsyToolkit.

PsyToolkit does not have any GUI for creating the questionnaires. To do this, it is required to use system format for the description of tests, which is entered on the corresponding page of the system. To create cognitive tests the internal language is used. This language contains a number of command and looks like programming language.

PsyToolkit allows to run own questionnaires and cognitive tests on mobile devices (in web browser), but it should be noticed that response time of touch screens is less accurate than the keyboard.

It is worth to pay attantion that it is prohibited to use PsyToolkit for commercial purposes or business unless direct consent has been obtained from the project author. Students can use PsyToolkit for study and own researches without additional extensions.

Characteristics of PsyToolkit:

- **Web-site–**http://www.psytoolkit.org/
- **License–**web version (Creative Commons Attribution-NonCommercial-ShareAlike 3.0), Linux (GNU GPL 3), commercial use prohibiled, free for students
- **Platform–**web service (SaaS), desktop (Linux)
- **Country of production–**no information
- **Functional purpose–**tool for cognitive-psychological experiments and questionnaires
- **Test types–**questionnaires, cognitive tests
- **Data storage resource–**on service servers (SaaS storage), locally (for Linux version)
- **Country of data storage–**France (Strasbourg)
- **Personal data storage features–**data is collected and stored on service servers (only for SaaS version), registration of researchers is necessary, there is an opportunity to remove account, the anonymity of the data collected depends on the researcher
- **Techniques of tests creation–**own test format (questionnaires), scripts programming (cognitive tests)
- **Test playing–**yes
- **Programming knowledge–**required for the use of questionnaire format and cognitive tests development
- **Complexity–**registration is required, interface looks outdated, easy use of prepared tests, difficult to create questionnaires and cognitive tests (required to develop sctipts)
- **Tools for data analyses–**no
- **Opportunities for data export–**text as output, export to .csv format
- **Integration tools–**Amazon Mechanical Turk, Qualtrix, Sona
- **Solutions relevance–**supported, the design is not updated and looks outdated
- **Solutions features–**there is library of experiments, demo tests, complete documentation, there is YouTube channel with training videos

G. 1ka

1ka is free web service (SaaS) of open source, that allows creation and online research completion. Web service includes questionnaire constructor with a large number of functions and opportunities. Interface of application is presented on Fig. 7.

1ka allows to create complex questionnaires that can include conditioning, blocks, loops and about 30 different types of questions. 1ka allows to send invitations and reminders to an unlimited number of potential respondents to participate in questionnaires and also allows to manage the database of respondents. In addition, it allows the user not only to use the public library of questionnaires, but also to create his own. 1ka includes embedded system of troubleshooting system for the created questionnaires and provides recommendations based on the methodological guidelines of the most common errors.

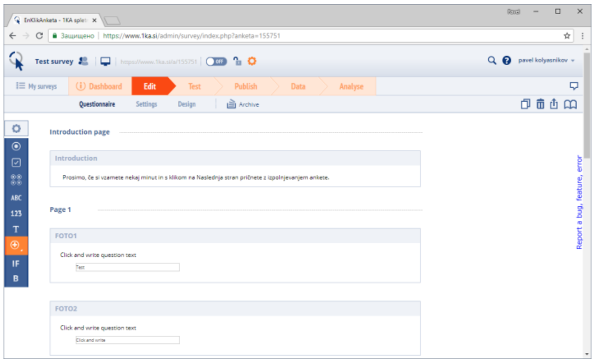
Characteristics of 1ka:

- **Web-site–**https://www.1ka.si/
- **License–**free (open source)
- **Platform–**web service (SaaS), mobile (Android)
- **Country of production–**Slovenia
- **Functional purpose–**creation and completion of online researches based on questionnaires
- **Test types–**questionnaires
- **Data storage resource–**on service sesrvers (SaaS storage)
- **Country of data storage–**Slovenia, Ljubljana (Centre of Social Informatics)
- **Personal data storage features–**data is collected and stored on service servers (by installation on the own server data will stored locally), application complies with GDPR, researchers registration is necessary, there is an opportunity to remove account

- **Techniques of tests creation**–test constructor
- **Test playing**–yes (by copping questionnaire)
- **Programming knowledge**–not required
- **Complexity**–easy, required to register and start test creating (by using application as SaaS), interface is outdated and tangled, test constructor is convenient and there is many settings
- **Tools for data analyses**–no (basic statistical data analysis)
- **Opportunities for data export**–data export into .csv format
- **Integration tools**–API support, modules
- **Solutions relevance**–relevant, supported, popular among online questionnaire systems
- **Solutions features**–support of different design themes, public library of tests, installation on own server, uploading test results into paper format (.pdf and .rtf), complete documentation

H. Labvanced

Labvanced is commercial Internet platform for creation, completion and sharing of online experiments based on questionnaires and cognitive tests [15, 16]. Project was planned as “all-in-one” to combine different types of testing, methods, capabilities and data analysis into one single platform using the latest web technologies. Fig. 8 shows interface of test constructor.



https://www.1ka.si/admin/survey/index.php?anketa=155751

Introduction page

Introduction

Prosimo, če si vzamete nekaj minut in s klikom na Naslednja stran pričnete z izpolnjevanjem ankete.

Page 1

FOTO1

Click and write question text

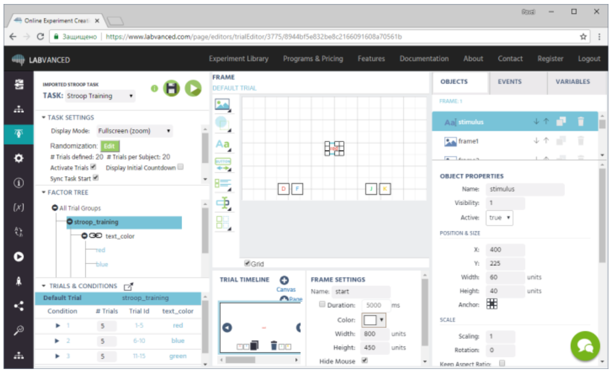
Test

FOTO2

Click and write question text

Click and write

Fig. 7. Interface of 1ka.



https://www.labvanced.com/page/editors/trialEditor/3775/8944bf5e832be8c2166091608a70561b

IMPORTED STROOP TASK
TASK: Stroop Training

TASK SETTINGS
Display Mode: Fullscreen (zoom)
Randomization: Edit
Trials defined: 20 # Trials per Subject: 20
Activate Trials ☑ Display Initial Countdown □
Sync Task Start ☑

FACTOR TREE
All Trial Groups
stroop_training
text_color
red
blue

TRIALS & CONDITIONS
Default Trial stroop_training

Condition	# Trials	Trial Id	text_color
1	5	1-5	red
2	5	6-10	blue
3	5	11-15	green

FRAME SETTINGS
Name: start
□ Duration: 5000 ms
Color:
Width: 800 units
Height: 450 units
Hide Mouse ☑

OBJECT PROPERTIES
Name: stimulus
Visibility: 1
Active: true

POSITION & SIZE
X: 400
Y: 225
Width: 60 units
Height: 40 units
Anchor:

SCALE
Scaling: 1
Rotation: 0
Keep Aspect Ratio: □

Fig. 8. Interface of Labvanced Test Constructor.

Researches in Labvanced consist of a task (an instance in the research), a block (a group of related tasks), a session (the session includes several blocks) and a group (used to define several groups of subjects in one research).

Labvanced allows to create the pages of different types: Canvas (a page of cognitive test with a free object replacement) and Page (a page of questionnaire). It allows to combine questionnaires and cognitive tests in one research. In Canvas, it is possible to add media objects (images, movies, audio content, eye tracking and sound recording). It should be noticed that it is not allowed to add images as object, but they can be added in text editor. Furthermore, Labvanced supports dynamic behavior due to the embedded event system (loops, conditions, callback, arrays, and mathematical expressions).

One of the interesting features of the platform is the support of real-time multiplayer experiments to study cooperative behavior, collaborative solutions etc. This feature requires two or more participants to simultaneously begin the same study.

Characteristics of Labvanced:

- **Web-site**–https://www.labvanced.com/
- **License**–chargeable, there is free tariff (1 active research, 10 uploads per month), open source planned for the future
- **Platform**–web service (SaaS)
- **Country of production**–Germany
- **Functional purpose**–online researches
- **Test types**–questionnaires, cognitive tests
- **Data storage resource**–on service servers (SaaS storage)
- **Country of data storage**–Germany
- **Personal data storage features**–data is collected and stored on service servers, required to register researchers (account removing and email changing are not possible), research participants are not required to register
- **Techniques of tests creation**–advanced constructor with support for both questionnaires and cognitive tests
- **Test playing**–yes (test coping in the system)
- **Programming knowledge**–not required
- **Complexity**–usage requires registration, the interface is thoughtful and looks good, increased learning curve due to the large number of features and settings, some actions are not intuitive
- **Tools for data analyses**–data review, filters, automated data analyses and visualization are planned in the future
- **Opportunities for data export**–data export into .csv and .json
- **Integration tools**–Amazon Mechanical Turk, API is planned in the future

- **Solutions relevance–**relevant, supported and updated
- **Solutions features–**combination of questionnaires and cognitive tests in one research, support of web cam for eye tracking and audio recording, support of cooperative researches (two and more participants), teamwork on one research (one person at a time), experiments can be compiled and loaded for offline research (premium users), there is a public library of experiments

I. Gorilla

Gorilla is a commercial web service, allowing to conduct online behavioral researches based on questionnaires and cognitive tests. For each type of test, GUI constructor is used and opportunities of visual programming shown in Fig. 9 are available. Gorilla also allows to develop some task with the use of JavaScript.

Web service Gorilla provides a large number of graphic constructor for creating experiments. All changes made are saved. The questionnaire constructor shows changes in real time, and the experiment constructor contains a graphic tree for visual programming of tasks. Gorilla includes different opportunities for development such as code editor, task builder, questionnaire builder. Furthermore, the integration with third-party libraries (i.e., JSPsych) is possible.

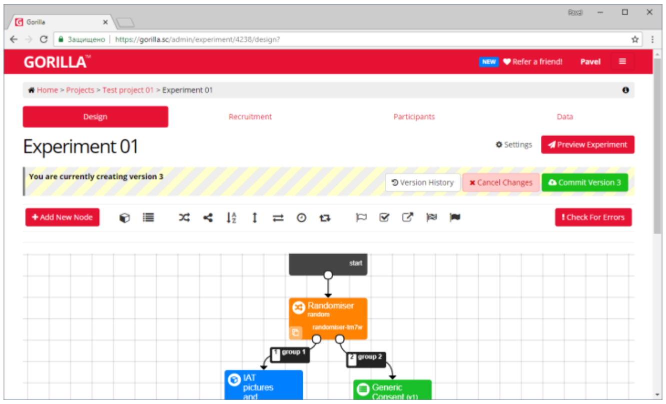
Gorilla supports different methods for participant invitation (email, link etc.). The application provides collaboration and sharing tools that allow users to share work with colleagues and co-workers.

Characteristics of Gorilla:

- **Web-site–**https://gorilla.sc/
- **License–**chargeable (charged for each respondent)
- **Platform–**web service (SaaS)
- **Country of production–**Republic of Seychelles
- **Functional purpose–**online behavior experiment completion
- **Test types–**questionnaires, cognitive tests
- **Data storage resource–**on the service servers (SaaS storage, Microsoft Azure)
- **Country of data storage–**Holland
- **Personal data storage features–**data is collected and stored on service servers, researchers registration required, there is an opportunity to remove account
- **Techniques of tests creation–**advanced graphic constructor
- **Test playing–**yes
- **Programming knowledge–**not required (only by use of scripts for functionality extension)
- **Complexity–**registration required, the interface is thoughtful and looks good, there are hints for new users, visual programming is used, the test constructor is overloaded in some places, there is a preview and good documentation
- **Tools for data analyses–**no
- **Opportunities for data export–**data export into .csv format
- **Integration tools–**integration with different systems (Prolific, Sona Systems, Amazon Mechanical Turk and Qualtrics), third-party libraries (i.e., JSPsych) and programming (JavaScript)
- **Solutions relevance–**relevant, supported and being updated
- **Solutions features–**creating an experiment in the form of visual programming, co-working on projects, saving versions of all the changes

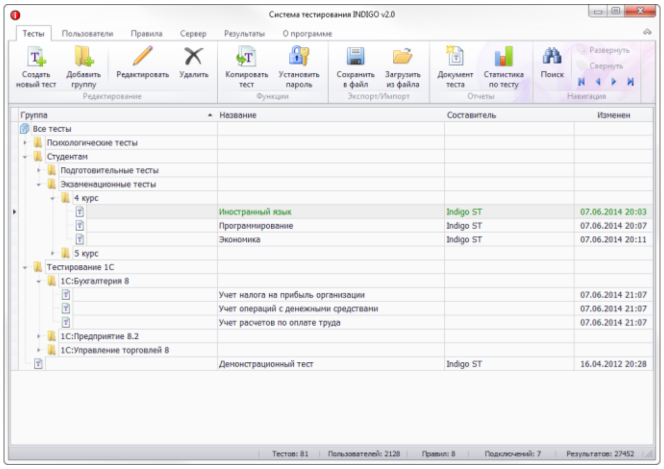
J. Indigo

Testing application Indigo is a tool of automated test and results processing. This tool is developed to solve a wide range of tasks: testing and student knowledge control, determining the professional level of employees, psychological testing, conducting questionnaires, organization of competitions and contests. Fig. 10 shows the interface of the administration panel.



You are currently creating version 3

Fig. 9. Interface of Gorilla.



Группа	Название	Составитель	Изменен
Все тесты			
Психологические тесты			
Студентам			
Подготовительные тесты			
Экзаменационные тесты			
4 курс			
	Иностранный язык	Indigo ST	07.06.2014 20:03
	Программирование	Indigo ST	07.06.2014 20:07
	Экономика	Indigo ST	07.06.2014 20:11
5 курс			
Тестирование 1С			
1С:Бухгалтерия 8			
	Учет налога на прибыль организации		07.06.2014 21:07
	Учет операций с денежными средствами		07.06.2014 21:07
	Учет расчетов по оплате труда		07.06.2014 21:07
1С:Предприятие 8.2			
1С:Управление торговлей 8			
	Демонстрационный тест	Indigo ST	16.04.2012 20:28

Тестов: 81 | Пользователей: 2128 | Правил: 8 | Подключений: 7 | Результатов: 27452

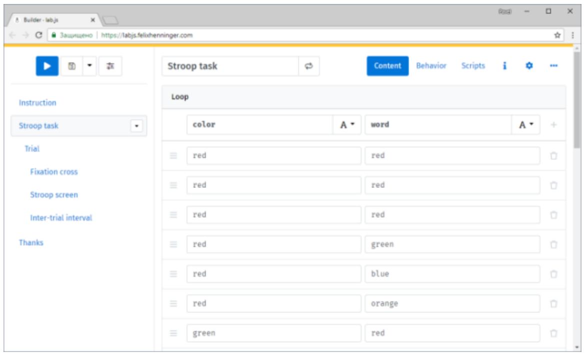
Fig. 10. Interface of Indigo Administration Panel.

The testing system is being installed on a single server using an installation package. The system can work both on an isolated computer and in a local network or via the Internet. Research participants work through web browsers, including on mobile devices.

One license allows working the testing system on one computer. The number of administrators is unlimited. The required maximum number of simultaneously tested users (number of connections) determines the cost of the license.

Characteristics of Indigo:

- **Web-site–**https://indigotech.ru/
- **License–**chargeable (depends on number of concurrent connections, permanent)
- **Platform–**client side is web browser, server and administration panel are Windows
- **Country of production–**Russian Federation
- **Functional purpose–**testing of students and staff, psychological testing, conducting questionnaires, organizing competitions and contests
- **Test types–**questionnaires
- **Data storage resource–**own local storage (installation on own server)
- **Country of data storage–**no
- **Personal data storage features–**data is collected and stored locally, researcher is responsible for its storing, anonymous testing is not provided (registration of respondents is required)
- **Techniques of tests creation–**test constructor
- **Test playing–**yes (test export and import in .itest format)
- **Programming knowledge–**not required
- **Complexity–**simple, it is enough to install the application using the installer and configure the system, the interface is thoughtful and looks beautiful
- **Tools for data analyses–**table of test results, statistics, diagrams are not realized
- **Opportunities for data export–**generating different reports and statistics, data export into .xls
- **Integration tools–**no
- **Solutions relevance–**relevant, being updated and used in a large number of organizations, schools and institutions
- **Solutions features–**includes a wide range of prepared tests, it is possible to order own configuration



Stroop task

color	word
red	red
red	red
red	red
red	green
red	blue
red	orange
green	red

Fig. 11. Interface of Test Constructor in Lab.js.

K. *lab.js*

lab.js is a tool allows to create any researches in browser in the field of social and cognitive science [17]. The researches can be created with the use of visual constructor or with the help of programming code [18]. Fig. 11 shows the interface of test constructor in lab.js.

lab.js was primarily focused on building experiments from scratch on HTML and JavaScript. The software solution is developed for people with programming skills. Lab.js includes a visual test constructor and JavaScript library.

The visual tool simplifies many aspects by research building. However, to use the constructor, it is needed to have basic programming skills and knowledge of HTML. The constructor includes various components: Canvas (visual placement of elements), Screen (page in HTML code), Sequence (shows components sequentially) and Loop (repeating component). In addition, full access to the base code is available. It will allow to adapt the research to any requirements.

The JavaScript library is the primary one used to build and run research lab.js. This allows to develop own researches without constructor from scratch.

The feature of the application is quite flexible possibilities of using a software solution: exporting an experiment for local use without a server, exporting an experiment with a server part in PHP, and also exporting to JATOS (in the zip archive format).

Characteristics of lab.js:

- **Web-site–**https://lab.js.org/
- **License–**open source (library of Apache License, constructor and interface of GNU Affero General Public License)
- **Platform–**web
- **Country of production–**Germany
- **Functional purpose–**tools for creation of social and cognitive tests

- **Test types–**cognitive tests
- **Data storage resource–**own local or remote storage
- **Country of data storage–**no
- **Personal data storage features–**data is collected and stored locally, researcher is responsible for its storing
- **Techniques of tests creation–**constructor (graphic interface), programming (HTML, CSS, JavaScript)
- **Test playing–**yes (experiments are saved in .json format)
- **Programming knowledge–**required (for basic understanding by using constructor and programming)
- **Complexity–**complex, test constructor contains some specific terms, the use of some features requires programming skills and reading the documentation
- **Tools for data analyses–**no
- **Opportunities for data export–**data export into .csv or .json format
- **Integration tools–**JATOS (import from lab.js), uploading to Netlify, integration with LimeSurvey, Qualtrics, SoSci
- **Solutions relevance–**relevant, the library is buing used and updated, but there is not any planning in the future
- **Solutions features–**there is a good flexibility due to the possibility of programming, supports the export of the test for local or online use (PHP, Netlify), there are training video materials.

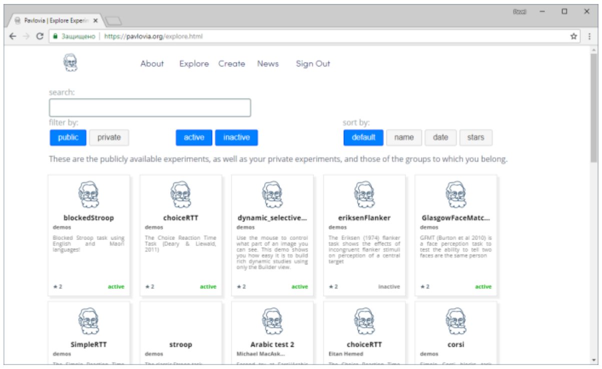
L. Pavlovia

Pavlovia is new service launched by PsychoPy developers in July 2018 [19]. The service is a community of researchers and a place for launch, exchange and study of experiments if the field of behavior science. Initially, the service was planned as a place for storing and launching researches created in PsychoPy. However, it was developed as an open solution that is aimed to support third-party tools such as jsPsych and lab.js. Fig. 12 shows a list of public and private prepared tests.

Pavlovia is not a service for research creation, but a service for storing, start and managing the experiments created with a help of PsychoPy (in visual tool PsychoPy Builder a test is being created, then it is presented on web page based on PsychoJS), jsPsych (high-level library on JavaScript), lab.js and other experiments, written on JavaScript.

All researches (projects) are stored in the Git version control system based on the free GitLab repository management system. Thus, the researcher needs to understand the principle of working with version control systems and a number of specific terms. Each project in GitLab is a research that will be displayed in Pavlovia (in an public or private list).

In Pavlovia there are demonstration tests that can be passed. It is possible to check their program code. It is worth noting that all the data obtained as a result of passing the test is stored in the same Git repository in the appropriate directory.



These are the publicly available experiments, as well as your private experiments, and those of the groups to which you belong.

blockedStroop
demos
Blocked Stroop task using English and Maori languages!
★ 2 active

choiceRTT
demos
The Choice Reaction Time Task (Deary & Liewald, 2011)
★ 2 active

dynamic_selective...
demos
Use the mouse to control what part of an image you can see. This demo shows you how easy it is to build rich dynamic studies using only the Builder view.
★ 2 active

eriksenFlanker
demos
The Eriksen (1974) flanker task shows the effects of incongruent flanker stimuli on perception of a central target
★ 2 inactive

GlasgowFaceMatc...
demos
GFMT (Burton et al 2010) is a face perception task to test the ability to tell two faces are the same person
★ 2 active

SimpleRTT
demos

stroop
demos

Arabic test 2
Michael MacAsk...

choiceRTT
Eitan Hemed

corsi
demos

Fig. 12. Window of Prepared Tests in Pavlovia.

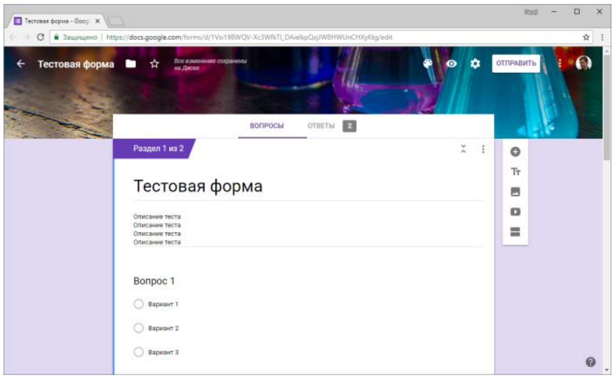
Characteristics of Pavlovia:

- **Web-site–**https://pavlovia.org/
- **License–**free
- **Platform–**web service
- **Country of production–**Great Britten (with support of University of Nottingham)
- **Functional purpose–**community of researchers and a place to start, exchange and study experiments in the field of behavioral sciences
- **Test types–**cognitive tests
- **Data storage resource–**on service servers
- **Country of data storage–**no information (domain is replaced in Nottingham, Great Britten)
- **Personal data storage features–**system complies with GDPR, authorization is required to create research (only e-mail is necessary), all data is not stored anywhere else by removing, the responsibility for data collection is borne by the researcher
- **Techniques of tests creation–**interface (PsychoPy Builder creates JavaScript with the use of PsychoJS library), programming (HTML, JavaScript, jsPsych, lab.js etc.)
- **Test playing–**yes (based on the git technology, there is an opportunity to move and copy, there is a revision history)
- **Programming knowledge–**required for test creation, git system knowledge is needed
- **Complexity–**difficult, because for the creation of tests, it is required to have a number of knowledge (programming and working with version control systems that programmers use)
- **Tools for data analyses–**no
- **Opportunities for data export–**data is collected and saved in .csv format
- **Integration tools–**no
- **Solutions relevance–**started in July 2018

- **Solutions features–**Git used, based on the Gitlab system, application can use PsychoPy Builder, jsPsych, lab.js and JavaScript for test creation, there is public research library.

M. Google Forms

Google Forms is online service for creation of feedback form creation, online testing and questionnaires. Tests can be created, edited and completed both on the computer and mobile device. Fig. 13 shows form editor in Google Forms.


Тестовая форма - Goog
Защищено
https://docs.google.com/forms/d/1Vsi198WQV-Xc3WfkTl_DAielkpQxjW8HWUnCHXyKtg/edit
Тестовая форма
Все изменения сохранены на Диске
ОТПРАВИТЬ
ВОПРОСЫ
ОТВЕТЫ
2
Раздел 1 из 2
Tт

Тестовая форма

Описание теста
Описание теста
Описание теста
Описание теста

Вопрос 1

○ Вариант 1

○ Вариант 2

○ Вариант 3

Fig. 13. Editor of Google Forms.

Google Forms contains one link for all research participants. Therefore, by turning on the option "Send the form no more than once", log in into Google account is required to participate in the questionnaires. It may also affect the anonymity of the data provided. At the beginning of the research, it is possible forcibly to ask for email addresses.

Google Forms provides a script editor in its language called Google Apps Script that is based on JavaScript.

Test results are being added to the table in real time. A respondent sends the answer that is saved in the table automated. Then it is possible to download the results in .csv format.

Characteristics of Google Forms:

- **Web-site–**https://www.google.com/forms/about/
- **License–**free
- **Platform–**web service (SaaS)
- **Country of production–**USA
- **Functional purpose–**creation of feedback form, online tests and questionnaires
- **Test types–**questionnaires
- **Data storage resource–**Google servers (SaaS)
- **Country of data storage–**no information provided (USA and disturbed servers all over the world)
- **Personal data storage features–**data is collected and stored on service servers, Google account required for test creation and sometimes for test completion, data collecting depends on the test
- **Techniques of tests creation–**visual test constructor
- **Test playing–**yes (by coping forms to Google Drive)
- **Programming knowledge–**not required
- **Complexity–**easy, all actions are performed in graphical mode with a preview, a small number of settings reduces the threshold for entry into the system, themes are supported
- **Tools for data analyses–**general summary of the answers (with diagrams), view of the results for each answer
- **Opportunities for data export–**data export into .csv format and Google Tables
- **Integration tools–**extensions, script editor (Google Apps Script)
- **Solutions relevance–**relevant, popular, used all over the world for questionnaires, testing and also for academic purposes
- **Solutions features–**teamwork on the creation, support of themes, editing the test after the start of the research, email notifications

Google Forms was chosen as the main and most well-known service for conducting online questionnaires. There is also a large number of similar services that provide similar functionality. They are not developed to conduct psychological researches, but may well be used for this purpose. It is also worth to note that these applications have a fairly simple data processing system.

III. Development Psychological Tools

Software analysis made it possible to formulate the development requirements for a specialized tool for longitudinal studies. Based on them, a Digital psychological platform was developed.

Digital psychological platform is a project started in 2017 and being developed [19, 20]. The tool specializes in the field of psychology: psychological testing, data collection and analysis, support for mass and longitudinal researches. Fig. 14 shows the interface of Digital psychological platform.

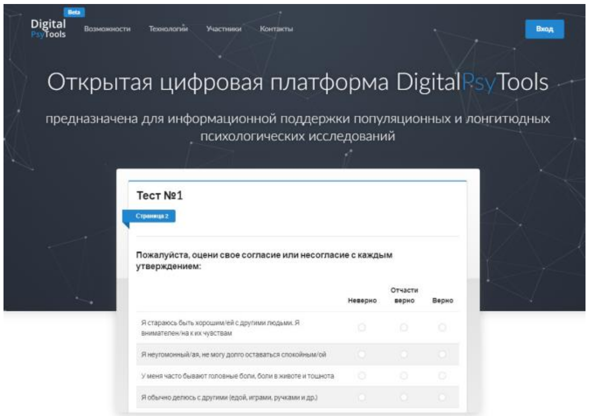
Currently, some of the planned features of the platform have not been implemented, and since the application is under development, it is early to speak of its effectiveness. Expected that the application will be used in the following areas: online and offline psychological researches conducting, mass researches at educational institutions, longitudinal surveys. The moderation is planned to verify the compliance of research ethical standards.

Characteristics of Digital psychological platform:

- **Web-site–**https://digitalpsytools.ru/
- **License–**free (registration requires approve from the service administrator)

- **Platform–**web-service (SaaS), desktop planned (Windows, Linux, MacOS) and mobile (Android, iOS)
- **Country of production–**Russian Federation
- **Functional purpose–**online questionnaires
- **Test types–**questionnaires, cognitive tests planned
- **Data storage resource–**on service servers
- **Country of data storage–**Russian Federation
- **Personal data storage features–**data is collected and stored on service servers, researchers registration is requires (confirmed by the administration of the service), anonymous testing is supported (an authorization mechanism is planned for a previously created set of logins/passwords without using personal data)
- **Techniques of tests creation–**test description format
- **Test playing–**yes (test export and import into .zip format)
- **Programming knowledge–**required for test creation
- **Complexity–**medium, the system interface requires a more detailed design, creating tests requires knowledge of the provided JSON format
- **Tools for data analyses–**output results table
- **Opportunities for data export–**data export into .csv and .json format
- **Integration tools–**API
- **Solutions relevance–**under development

Proposed an approach to conduct researches in the condition of an unstable and slow Internet connection, proposed a mechanism for ensuring reproducible research (transmission in unchanged form), and developed its own format of tests on based JSON.



Открытая цифровая платформа DigitalPsyTools

предназначена для информационной поддержки популяционных и лонгитюдных психологических исследований

Тест №1

Пожалуйста, оцени свое согласие или несогласие с каждым утверждением:

	Неверно	Отчасти верно	Верно
Я стараюсь быть хорошим/ей с другими людьми. Я внимателен/на к их чувствам	○	○	○
Я неугомонный/ая, не могу долго оставаться спокойным/ой	○	○	○
У меня часто бывают головные боли, боли в животе и тошнота	○	○	○
Я обычно делюсь с другими (едой, играми, ручками и др.)	○	○	○

Fig. 14. Interface of Digital Psychological Platform.

IV. Conclusion

This digital platform was used for project of the "Cross-cultural Longitudinal Analysis of Student Success" (CLASS). The aim of this longitudinal genetically informative project is the study of the factors and mechanisms of success in learning throughout the school years. Cross-cultural longitudinal study is aimed to study the whole period of schooling - from 1 to 11 grade. A longitudinal research involves 11 stages of measurements (each year of study - one stage of research). Schoolchildren from two states (the Russian Federation and the Kyrgyz Republic), who are similar in organization of the education system, but differ in language, culture, structural and institutional characteristics of the society, take part in the study. The research sample includes 1300 schoolchildren from Russia and 1700 from Kyrgyzstan (including schoolchildren with a native Kyrgyz language and schoolchildren with a native Russian language). In total, 3,000 schoolchildren will take part in the study, the age range of the participants covers the entire period of schooling (from 6.5 to 18 years).

Acknowledgment

This study was supported by the grant from the Russian Science Foundation № 17-78-30028.

References

[1] R. M. Epstein, "Assessment in medical education," New England journal of medicine, vol. 356, no. 4, pp. 387-396, 2007.

[2] J. L. Jolly, Historical Perspectives: Lewis Terman: Genetic Study of Genius—Elementary School Students, Gifted Child Today, vol. 31, no. 1, pp. 27-33, 2008.

[3] J. Peirce & M. MacAskill, "Building Experiments in PsychoPy". SAGE. 2018

[4] H. Bhin, Y. Lim, S. Park, & J Choi, "Automated psychophysical personality data acquisition system for human-robot interaction," In 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), IEEE, pp. 159-160, 2017

[5] PsychoJS [Online].Available: http://psychopy.org›online/psychojsCode.html

[6] T. Yamauchi, A. Leontyev, & M. Wolfe, "Choice reaching trajectory analysis as essential behavioral measures for psychological science," Insights in Psychology, vol. 1, no. 4, 2017

[7] S. Mathôt, D. Schreij, & J. Theeuwes, "OpenSesame: An open-source, graphical experiment builder for the social sciences," Behavior research methods. vol. 44, no. 2, pp. 314-324, 2012

[8] S. Rajananda, H. Lau, & B. Odegaard, "A Random-Dot Kinematogram for Web-Based Vision Research," Journal of Open Research Software, vol. 6, no. 1, 2012

[9] S. T. Mueller, & B. J. Piper, "The psychology experiment building language (PEBL) and PEBL test battery," Journal of neuroscience methods, vol. 222, pp. 250-259, 2014

[10] M. S. Clair, "Trail Making Test: Comparison of the PEBL and iPAD Versions," Archives of Physical Medicine and Rehabilitation, vol. 98, no. 10, pp. e121-e122, 2017

[11] J. Katona, & A. Kovari, "The Evaluation of BCI and PEBL-based Attention Tests," Acta Polytechnica Hungarica, vol. 15, no. 3, pp. 225–249, 2018.

[12] K. Lange, S. Kühn, & E. Filevich, "Just Another Tool for Online Studies"(JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies," PloS one. vol. 10, no. 6, pp. e0130834, 2015.

[13] G. Stoet, “PsyToolkit: A software package for programming psychological experiments using Linux,” Behavior Research Methods, vol. 42, no. 4, pp. 1096-1104, 2010.

[14] G. Stoet, “PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments,” Teaching of Psychology, vol. 44, no. 1, pp. 24-31, 2017.

[15] H. Finger, C. Goeke, D. Diekamp, K. Standvoß, &, P. König, “LabVanced: A Unified JavaScript Framework for Online Studies,” In 2017 International Conference on Computational Social Science IC2S. Cologne, 2016.

[16] E. Munoz-de-Escalon, & J. Canas, “Online measuring of available resources,” In H-Workload 2017: The 1st international symposium on human mental workload,–Dublin Institute of Technology, 2017. doi:10.21427/D7DK96

[17] F. Henninger, U. K. Mertens, Y. Shevchenko, & B. E. Hilbig, “lab.js: Browser-based behavioral research”, 2017. doi: 10.5281/zenodo.597045.

[18] P. Garaizar, & U. D. Reips, “Best practices: Two Web-browser-based methods for stimulus presentation in behavioral experiments with high-resolution timing requirements,” Behavior research methods, pp. 1-13, 2018. doi:org/10.3758/s13428-018-1126-4

[19] E. Nikulchev, D. Ilin, P. Kolyasnikov, V. Ismatullina, & I. Zakharov, “Development of a Common Format of Questionnaire Tests for a Web-based Platform of Population and Experimental Psychological Research”, ITM Web of Conferences, vol. 18, pp. 04004, 2018.

[20] E. Nikulchev, D. Ilin, P. Kolyasnikov, V. Belov, I. Zakharov, & S. Malykh, “Programming technologies for the development of web-based platform for digital psychological tools”, International Journal of Advanced Computer Science and Applications, vol. 9, no. 8, pp. 34-45, 2018.

An Ontological Model for Generating Complete, Form-based, Business Web Applications

Daniel Strmečki[1], Ivan Magdalenić[2]
Faculty or Organization and Informatics, University of Zagreb, Varaždin, Croatia

***Abstract*—This paper presents an ontological model for specifying and automatically generating complete business Web applications. First, a modular and expandable ontological model for specifying form-based, business Web applications is developed and presented. Next, the technology used for transforming the ontological specification to Java executable code is explained. Finally, the results of applying the proposed model for specifying and generating an order management application are presented. Results showed that the application of an ontological model in a generative programming approach increases the level of abstraction. This approach is especially suitable for development of software families, where similar features are reused in multiple products/applications.**

Keywords—Ontology; model; generative; automatic; programming; development

I. Introduction

Software reuse is an important area of software engineering discipline, as it can increase software productivity and quality. However, factors like deadlines, budget, technology, architecture and the level of knowledge and experience also need to be taken into account. Raising the abstraction level is the most commonly used approach to increase software reuse. By encapsulating knowledge about lower level operations, developers and engineers can think in terms of higher level concepts, thus saving effort and time [1], [2].

Automatic programming is a software engineering discipline that automates the lower level process. Its main goal is to enable developers to operate on higher abstraction levels by making machines do parts of the programming work. Even after more than 60 years since its appearance, there are still no universal solutions for software development automation. However, it must be acknowledged that the discipline has dramatically influenced on improvement of software developers' productivity. A high level of automation in software development can nowadays be achieved in some domains, but the dream of enabling general-purpose, full development automation still remains unrealized [3].

Generative programming is a sub discipline of automatic programming that uses generators to facilitate the process of application development. A generator is piece of software that takes a higher-lever specification of another piece of software and produces its implementation [4]. A domain model is used to provide mapping between the problem space and solution space. Problem space in generative programming refers to a set of software features to be implemented, while solution space implies the implementation abstractions contained in the specification. A generator maps the two spaces by using a specification to yield the corresponding implementation [5].

Ontologies were initially applied in software development only to store data and their semantic meaning, but nowadays they are used in all phases of software development lifecycle. Even a new discipline Ontology-Driven Software Engineering arose, where ontologies are used to perform a majority of operations in software development [6]. Several authors found ontologies suitable and helpful in performing tasks like description of specification documents, formal representation of requirements, semantic description of services/components, domain model generation, test cases generation and executable code generation [7]–[16]. In this paper we present and apply an ontological model for generating business Web applications. We applied ontologies for both modeling and specifying a complete, form-based application. The ontological specification is then used as an input to code generators, which can produce fully functional Web applications.

This paper is organized as follows. Section 2 describes the related work. It describes and discusses models and frameworks used for the same purpose as the one presented here. Section 3 introduces a modular ontological model for modelling and generating business Web applications. Section 4 describes how the ontological model and its specification are used for generating complete, form-based, applications. Section 5 provides an example application of the developed ontological model. An order management business Web application is ontologically specified and generated. Section 6 presents our conclusions.

II. Related Work

In their approach named Ontology Driven Architecture for Software Engineering authors Bossche et al. proposed the usage of ontologies for forming a knowledge continuum between business and IT. First, the business representatives work closely with domain modelling experts to build a formal business model. This enables the business to formalize their specification and forces them to make requirements explicit. Next, once the ontology is formed, a transformation from ontology to source code is done automatically. For this purpose, authors used a platform based on a strongly typed logic programming language Mercury and a set of tools and libraries called Hedwig [17]. They pointed out a number of advantages brought by the use of ontologies in automated programming, but it remains unclear what ontologies were used and how were they mapped to source code generators.

Authors Tang et al. developed a Web Information System auto-construction Environment, which is also a platform for automatic source code generation based on ontologies. Their platform uses a predefined ontology to specify user requirements and provides graphical tools for ontology construction and drawing user interfaces. It is based on three tools: builder, mapper and generator. The builder tool helps users construct the domain and behavior ontologies. The mapper tool provides methods to automatically transform behavior operations to backend database access code. The generator tool, as the name suggests, generates the source code from ontologies [18]. Since the platform generates program code automatically from predefined ontologies, it is not possible to add new ontological or user interface elements without extending the platform itself.

Semantic Web Builder is an agile development platform for Web application development, proposed by authors Solis et al. In their platform requirements are modeled using ontologies and the infrastructure is then automatically generated. The source code is however generated in two layers. A base layer is automatically generated and should never be modified. For implementation of specific functionality, an extended layer is also provided [19]. This presents the main disadvantage of their approach, as custom code needs to be developed manually for every specific functionality.

In this paper we will present an approach for generating complete Web applications from ontologies. We will explain how the ontological model fits into a source code generation platform, built in Java using only publicly available tools and frameworks. We will provide public access to the ontological model and the source code of the generators we have developed. This will make it possible for anyone to extend the model and generators, with their custom tweaks or new functionality. The goal of the proposed platform is to be able to model and generate any new or specific features trough ontologies, so there is no custom code and all features can be easily reused in future work. We rely on ontologies for modeling complete systems, generate fully functional applications and avoid repetitive operations, including any code or specification copy-pasting. The presented approach is limited to working with form-based Web application and relational databases.

III. Ontological Model

Given that we wanted to achieve a high level of reusability and expandability of the initially created ontological model, we decided to take a modular approach in the model creation. We defined five initial ontologies, where each is focused on one subdomain: 1) Requirements ontology, 2) Repository ontology, 3) Web forms ontology, 4) Web components ontology and 5) User interface ontology. Requirements ontology contains classes and attributes defining the project, client and requirements, including epics, user stories and acceptance criteria. Repository ontology contains classes and attributes defining a relational database, its tables and columns. In the initial version, only relational databases are supported, but other classes supporting different types of repositories could be developed in the future. Web forms ontology defines all forms available to the user through the application, as well as how forms are grouped together and displayed in the main menu. Web components ontology contains definitions of components and listeners. Components are further divided into input fields, action buttons, pagination tables, etc. The input fields are divided into the standard input field types we use on modern Web forms, such as text input fields, numeric input fields, select fields, combo fields, date input fields, etc. User interface ontology contains classes that define layout attributes and positioning of the components on the screen. *Fig. 1* shows the connections between the mentioned ontologies and how they fit into a complete ontological model for specifying form-based Web applications. The hierarchy of the complete ontological model is shown in *Fig. 2*. The complete model consists of 2300+ axioms, 30+ classes and 80+ types of properties, from which 37% are object properties and the rest are data properties.

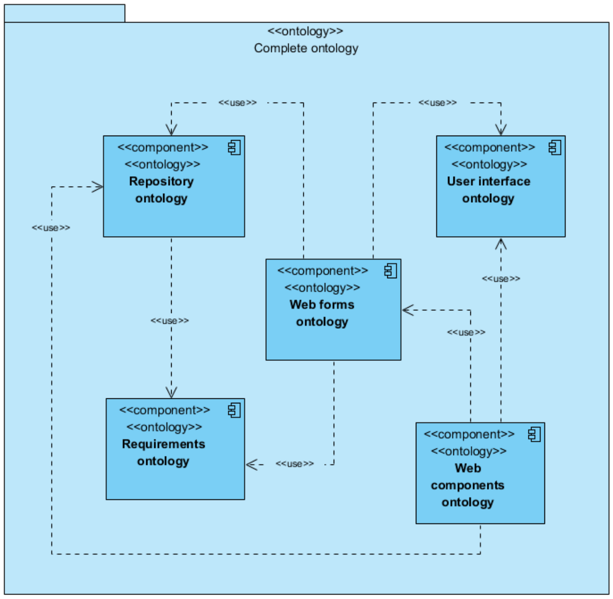


Fig. 1. Ontological Model Components.

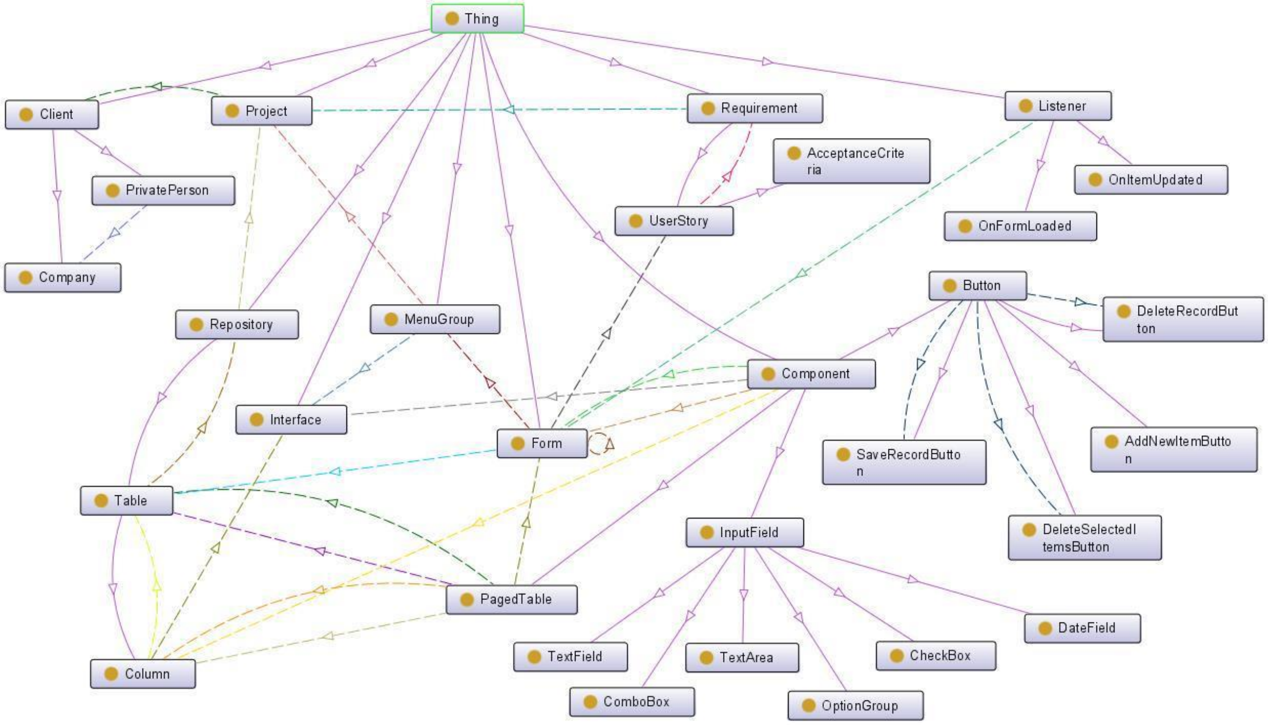


Fig. 2. Complete Ontological Model.

IV. Code Generation

The language of choice for development of code generators is Java, simply because it is a very popular object-oriented language, in which the author is working professionally for several years. Java also has a very good support for working with code templates, Web components and ontologies. Apache Jena is the framework we used for fetching the ontological model structure and querying the ontology in order to retrieve specification. We also used Apache Velocity framework for templating Java and HTML/CSS code. Velocity allows developers to use a simple template language with access to reference objects defined in the Java code. Next, we used Vaadin Web framework which allowed us to use a large number of free, out-of-the-box Web components. Vaadin uses a component-based approach for rapid development of user interfaces for Java Web applications. This is very convenient, because we want to focus on development of code generators, not user interface components. Finally, we used Hibernate framework as our object-relational mapping framework of choice. Hibernate enables us to define database entities as Java classes and is able to generate the database schema automatically.

The purpose of the ontological model and specification is to provide an input for code generation. As presented in *Fig. 3*, a code generator first fetches the ontological model vocabulary and then the specific application specification that needs to be generated. The ontological model vocabulary is represented in the generators source code by a single Java class for each ontology. These classes are automatically generated using a tool called Schemagen, provided with Apache Jena framework. We use these generated classes to retrieve data from the ontological model. Application specification data is retrieved from ontology by running SPARQL queries from generators Java code. Code generators map the specification of forms, components and database entities to appropriate Velocity code templates. This process results in generated files/classes of the final application. Once all the files are generated, the generation process results in a complete, fully functional Java Web application/project. The application code can be compiled and installed on an appropriate Java web server like Jetty or Tomcat. Besides the Web server, it is also necessary to provide a connection from the server to a relational database. We used MySQL database in our example application and Hibernate helped us to automatically create the database schema. Once the infrastructure is set up and the application is successfully installed, users can access the generated application forms through their Web browser.

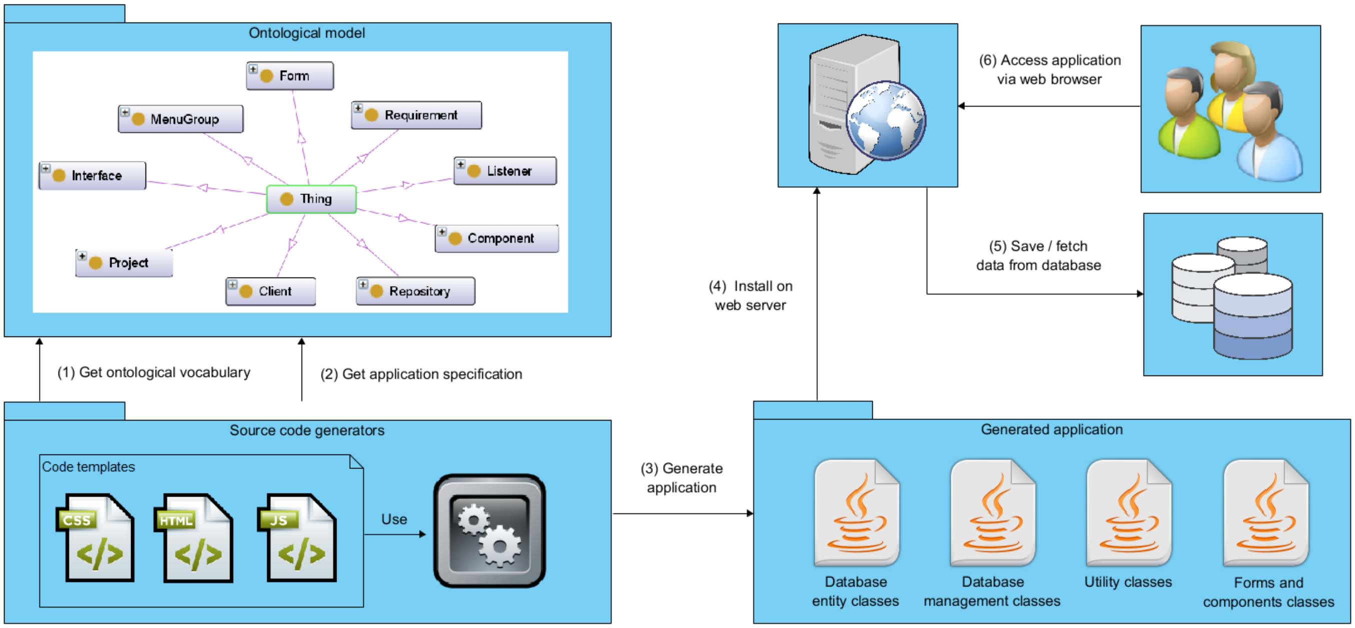


Fig. 3. Using an Ontological Model to Generate Executable Code.

V. Application

As an example application we ontologically specified and generated an order management application. This sample application contains a total of 18 Web forms including forms for managing addresses, articles, contacts, states, references, Incoterms, legal persons, private persons, orders and tax schemas. It is important to note that these are not only simple CRUD forms and that some of them contain complex relations. For example, it is possible to add items to existing orders and connect them to multiple referenced documents. When adding new items the total tax and payable amount get updated automatically. The generated forms contain almost 7.000 lines of Java code. The complete generated project resulted in 52 generated files, 10.000+ lines of code, 5500+ statements, 600+ methods and 70+ classes. Let's now analyses the ontological specification, based on the presented model, which was used to generate the application. In order to analyze only the specification part of the ontology, we exported the ontology in RDF/XML format and extracted only named individuals. From the resulting file we removed all blank lines and comments, so that each line represents either a named individual or a property. The ontological specification contains less than 2000 lines, 260+ instances and 1100+ attributes. The results show that we were able to generate more than 5 times more executable code than we specified using the ontological model. This means that we have successfully raised the development abstraction level. Also, if we were to develop a new application, with a similar set of features, we would be able to reuse the complete ontological model and even parts of the specification. Development efforts would be required only to support new features, which are not already present in the model.

The ontological model, source code of the code generators and the example application are available on this link: http://gpml.foi.hr/DanielStrmecki/.

VI. Conclusion

In this paper we presented an ontological model for specifying and automatically generating complete, business Web applications. In our approach we wanted to avoid the mix of generated and custom code, so all new features should be built in the existing model and full application generated on a single click. This approach enables possible future reuse of all developed features. In addition, we provided free, public access to the model and source code of the generators. The ontologically-enabled, generative programming approach presented here is especially suitable for developing software product families, with a significant set of common features. Since development and maintenance of the ontological model and code generators requires additional resource investments, this approach becomes feasible only if a significant amount of features are reused on multiple products/projects. As connections in the ontological specification are defined via object properties, even the same specifications can be reused between products. For example, a database entity for handling addresses, specified for one product, can be connected to another project instance simply by adding a new object property. In this way, we are able to reuse the model, code generators and the specification with minimal modeling effort. Development efforts are only need when introducing new features to the model or when debugging any of the currently available ones. Our approach reduces the number of repetitive programing tasks, enables development of software product families on high abstraction level and with high level of reusability.

Disclaimer

This paper publishes, in English, parts of the PhD thesis Framework for Business Web Application Families Development Using an Ontological Model and Source Code Generators by Daniel Strmečki, mentored by Ivan Magdalenić. The original work is published in Croatian, in accordance with the PhD study program on Faculty of organization and informatics, University in Zagreb, Croatia. To download the original work in Croatian, please visit the Croatian Digital Dissertations Repository[1].

REFERENCES

[1] X. Nianfang, Y. Xiaohui, and L. Xinke, "Software Components Description Based on Ontology," in 2010 Second International Conference on Computer Modeling and Simulation, 2010, vol. 4, pp. 423–426.

[2] E. Visser, "WebDSL: A Case Study in Domain-Specific Language Engineering," Gener. Transform. Tech. Softw. Eng. II, vol. 5235, pp. 291–373, 2008.

[3] D. Strmečki, I. Magdalenić, and D. Radosević, "A Systematic Literature Review on the Application of Ontologies in Automatic Programming," Int. J. Softw. Eng. Knowl. Eng., vol. 28, no. 5, pp. 559–591, May 2018.

[4] K. Czarnecki and U. W. Eisenecker, Generative programming: methods, tools, and applications. ACM Press/Addison-Wesley Publishing Co., 2000.

[5] I. Magdalenić, D. Radošević, and T. Orehovački, "Autogenerator: Generation and execution of programming code on demand," Expert Syst. Appl., vol. 40, no. 8, pp. 2845–2857, 2013.

[6] A. J. Wiebe and C. W. Chan, "Ontology driven software engineering," in 2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2012, pp. 1–4.

[7] H. Happel and S. Seedorf, "Applications of Ontologies in Software Engineering," 2nd Int. Work. Semant. Web Enabled Softw. Eng. (SWESE 2006), pp. 1–14, 2006.

[8] C. Calero, F. Ruiz, and M. Piattini, Ontologies for Software Engineering and Software Technology. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[9] F. Bartolo Espiritu, A. Sanchez Lopez, and L. J. Calva Rosales, "Towards an improvement of software development process based on software architecture, model driven architecture and ontologies," Int. Conf. Electron. Commun. Comput., pp. 118–126, 2014.

[10] D. Gašević, N. Kaviani, and M. Milanović, "Ontologies and Software Engineering," in Handbook on Ontologies, no. January 2016, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 593–615.

[11] E. K. Karatas, B. Iyidir, and A. Birturk, "Ontology-Based Software Requirements Reuse: Case Study in Fire Control Software Product Line Domain," in 2014 IEEE International Conference on Data Mining Workshop, 2014, pp. 832–839.

[12] A. S. Andreou and E. Papatheocharous, "Automatic Matching of Software Component Requirements using Semi-formal Specifications and a CBSE Ontology," Eval. Nov. Approaches to Softw. Eng., 2015.

[13] S. Il Kim and H. S. Kim, "Ontology-based open API composition method for automatic mash-up service generation," in 2016 International Conference on Information Networking (ICOIN), 2016, pp. 351–356.

[14] H. A. Duran-Limon, C. A. Garcia-Rios, F. E. Castillo-Barrera, and R. Capilla, "An Ontology-Based Product Architecture Derivation Approach," IEEE Trans. Softw. Eng., vol. 41, no. 12, pp. 1153–1168, Dec. 2015.

[15] R. Sinha, C. Pang, G. S. Martinez, J. Kuronen, and V. Vyatkin, "Requirements-Aided Automatic Test Case Generation for Industrial Cyber-physical Systems," Eng. Complex Comput. Syst. (ICECCS), 2015 20th Int. Conf., pp. 198–201, 2015.

[16] D. Toti and M. Rinelli, "Semi-automatic Generation of an Object-Oriented API Framework over Semantic Repositories," in 2015 International Conference on Intelligent Networking and Collaborative Systems, 2015, pp. 446–449.

[17] M. Vanden Bossche, P. Ross, I. MacLarty, B. Van Nuffelen, and N. Pelov, "Ontology driven software engineering for real life applications," Third Int'l Work. Semant. Web Enabled Softw. Eng, pp. 1–5, 2007.

[18] L. Tang et al., "WISE: A Prototype for Ontology Driven Development of Web Information Systems," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3841 LNCS, no. 60473072, 2006, pp. 1163–1167.

[19] J. Solis, H. Pacheco, K. Najera, and H. Estrada, "A MDE Framework for semi-automatic development of web applications," Model. 2013 - Proc. 1st Int. Conf. Model. Eng. Softw. Dev., pp. 241–246, 2013.

[1] https://dr.nsk.hr/en/islandora/object/foi%3A3579

Sound user Interface with Touch Panel for Data and Information Expression and its Application to Meteorological Data Representation

Kohei Arai
Graduate School of Science and Engineering, Saga University, Saga City, Japan

***Abstract*—Sound User Interface (SUI) with touch panel for representation of quantitative data and information together with its application to meteorological data representation is proposed. The proposed SUI is not a merely ear-con. Through experiments, it is found that the proposed SUI combined with visual perception makes meteorologist to understand meteorological data intuitively and is much understandable than ever in a comprehensive manner. It is also useful to hear "images" in particular, for blind person.**

***Keywords*—Audible; meteorology; remote sensing satellite; musical scale; multi-layered data; SUI**

I. Introduction

Recently, new user Interface: UI is invented and developed. Tangible Interface (TI)[1] is one of those. TI allows users touch data or information. It differs from touch interface. Touch interface allows caution/warning/awareness signs and mouse operations with touching "Touch Screen"[2]. On the other hand, the current Sound Unser Interface: SUI allows caution/warning/awareness signs and mouse operations no more than that. New SUI proposed here which allows not only caution/warning/awareness signs and mouse operations but also hear/listen data and information is my concern.

In the Apple Watch Series 3 and 4[3] released in September 2017 and quite recently, respectively, the voice command by Siri [4] has been further strengthened. Sound-based communication may become more important than ever. Recently, we have been surrounded by enormous amount of information which has not existed so far. Originally, humans acquire information by using various information about surroundings using five senses, but most information on current information equipment is displayed by visual media. However, in recent years, users' experiences in applications and Web services attracted attention in the user interface, UI of SUI which makes use of non-verbal sounds in addition to visual UI plus sound, especially voice.

One major problem emerges as the downsizing of equipment such as smart phones and wearable devices is accelerated. It means that there is a limit to the size of the visual display and the purpose of keeping staring at the display gradually diminishes. Due to miniaturization of devices, information display means to replace visual information has come to be desired. Actually, many of us have heard the sound as a UI. For example, on desktops such as Windows and MacOS, if you empty the trashcan, you will hear the "Crash sound". If this sound does not occur, the user does not know whether the trash can really be empty, once you open the trash can. However, as the sound is steadily generated, the user can be confident that the action of emptying the trash box has been completed.

The communication application such as LINE [5] and Facebook Messenger[6] also implements a function to notify that a message has been received by sound. This allows you to confirm that the message arrived instantaneously by picking up the sound even if the user is doing another work.

SUI has some advantages over visual UI, so it is important to use it effectively. Here let's examine the features of SUI and the three elements necessary to design SUI. The SUI is to convey some information, and there are messages (information) to be displayed there.

Sonification system for representation of satellite remote sensing data is proposed [1] together with sonification method for representation of multi-dimensional meteorological data derived from Earth observation satellites [2]. Recently, method for audible representation of meteorological data derived from remote sensing satellites is proposed [3]. Next section describes some related research together with the proposed method followed by experiments with atmospheric sounder data derived from remote sensing satellites as an example. Finally, some concluding remarks and discussions are described.

II. Related Research

SUI[7] is not so popular in the user interface research field. Adequate context for interpreting audibles of data is not so easy [4]. Also, meaning is many audible attempts are coded from scratch [5]. The opto-phone which consists of selenium photo-sensors to detect black print and convert it into an audible output was invented [6]. The first experiment of the transmission of information via auditory display was

[1] https://en.wikipedia.org/wiki/Tangible_user_interface

[2] https://en.wikipedia.org/wiki/Touchscreen

[3] https://www.apple.com/jp/apple-watch-series-4/?afid=p238%7CLoYxmRFF-dc_mtid_2092567642642&cid=wwa-jp-kwyh-watch

[4] https://ja.wikipedia.org/wiki/Siri

[5] https://line.me/ja/

[6] https://www.facebook.com/messenger

[7] http://sounduserinterface.org/sui/

published [7]. The "Auditory Data Inspection (ADI)" is proposed and reported in the technical memorandum [8]. The effectiveness of the ADI has not been done yet [9]. Another audible system, so called "Pulse Oximeter" which allows audible oxygen concentration of blood was reported [10]. International Community for Auditory Display (ICAD)[8] has been established and its conferences have been conducted [11]. There are some interactive audible techniques [12-14]. Model-Based Audible, Parameter Mapping Audible, and Stream-Based Audible are identified as difficulties [15].

SUI proposed here is the interface with touch panel for representation of quantitative data and information together with its application to meteorological data representation. The proposed SUI is a brand new method which allows hearing the satellite data as well as meteorological data. Through experiments, it is found that the proposed SUI combined with visual perception makes meteorologist to understand meteorological data intuitively and is much understandable than ever in a comprehensive manner. It is also useful to hear "images" in particular, for blind person.

III. Proposed Method

A. Design Concept of the Proposed SUI

These sound data are created by using "Sakura" software tool[9] (Free open source software for sound representation). It provides sound sources with the different music instrument types, musical scale, rhythm, loudness, harmony, etc. with text representations.

B. Detailes of Sakura

"Sakura" allows easily compose with" katakana" character such as "ドレミファソラシド: Doremifa solasido" representing the scale. In the built-in text editor, it is possible to quickly input symbols such as sharp, flat, octave raising and lowering from the list. Musical Instrument Digital Interface[10] (MIDI) type of output files can be created together with Windows Media Player[11] (WMP) type files. After composing, it is possible to play with MIDI, save it as a text file or MIDI file. It is possible to also play songs created with other text editors or songs sent by email from friends. In addition to the katakana scale, it also supports MML[12] which is widely used in computer music.

Since it has a script function, it can also extend its functions such as delicate musical expression and algorithmic composition. It is possible to also listen to lots of songs made by users on the author WEB page. For example, it is possible to play a tulip with "Dreme Remy Mile Dre Mile". Sakura is developed as a foundation of Music Macro Language [13] (MML). MML is a musical notation language devised to describe the Back Ground Music (BGM) of the game by expressing Doremifa solasi as "cdefgab".

Also, by writing a simple script in the song, there is also the merit that it is possible to expand its function. If you use Sakura, it is possible to make music of any genre from classical to pop, finally to experimental music, with a sense of word processor.

Sakura has a scripting function that allows you to use control structures such as If and While. If you enjoy music all the way to a senior class, try using the script, algorithm composition! Also, it is the real pleasure of Sakura that it is possible to enjoy music of the source like maniacal, such as concept music, modern music, experimental music and random number music.

Fig. 1 shows the initial displayed screen shot. All the available functions are aligned at the top row on the other hand, operation (or functional) menus are aligned at the left column.

Fig. 2 shows the operational display divisions. There are six divisions, text editor, keyboard input, enter staff notes, useful tab (functional menu), message display, and play monitor.

Detailed operations are as follows:

1) *Text editor:* Write performance information such as "Doremi" in the text editor.

2) *Button for stop the play:* Press the playback button.

3) *Message display:* If there is a mistake in the performance information, the message will be displayed in.

4) *Play monitor:* The performance monitor is displayed and the performance starts.

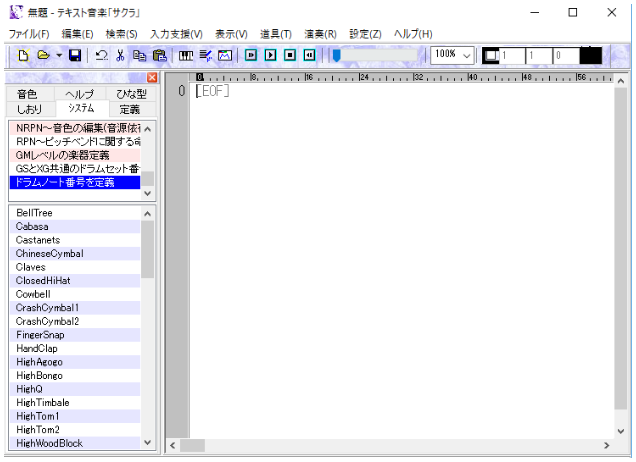


Fig. 1. Initial Display of Screen Shot.

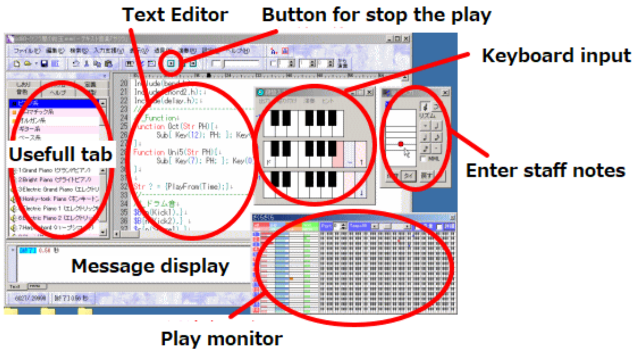


Fig. 2. Display Screen Divisions.

[8] http://www.icad.org/
[9] https://forest.watch.impress.co.jp/library/software/sakura/
[10] https://ja.wikipedia.org/wiki/MIDI
[11] https://ja.wikipedia.org/wiki/Windows_Media_Player
[12] https://ja.wikipedia.org/wiki/MMI
[13] https://ja.wikipedia.org/wiki/Music_Macro_Language

It is a colorful performance monitor, parameters per channel on the left side. The keyboard on the right side is colored with the key being pressed. It is possible to customize the color. When you click the keyboard or staff score with the mouse, the performance information of Doremi is inserted in the editor. In case of staff input, to enter sharp, press the [SHIFT] key.

1) Message tab: Messages, help, etc. are displayed.

2) Keyboard input: It is possible to insert Doremi text by clicking the keyboard indicated by the picture.

3) Score entry: It is possible to insert Doremi text by clicking on the staff.

4) Convenient tab: It is possible to watch the tone and insert instructions.

5) Bookmark: It is possible to set bookmarks in any portion of the editor.

By using simple bookmarks, it is possible to set bookmarks at arbitrary points (up to 10 places) and instantly return to that place. As for the method of setting simple bookmarks, when you right-click on the editor, the menu "register bookmark" appears, so it is possible to set bookmarks by selecting the registration number. It also supports shortcuts (same as Borland Delphi[14]), with Shift + Ctrl + Numbers, it is possible to go back to the bookmark with Bookmark Registration and Ctrl + Numbers.

The bookmark function is located on the command insertion tab, giving a name to arbitrary parts of the editor, so that it is possible to jump to that part immediately afterwards. To create a bookmark with the bookmark function, write a comment such as "// _ name" at the beginning of the line of the editor, in the editor, or right click on the editor, it comes out "bookmark It is possible to also do it using "Register".

The circulation function of song bulletin board; it is possible to quickly download and listen to user-created songs. Text music "Sakura" has been published on the Internet since November 5, 1999. The number of users has also increased. And, above all, many wonderful users are using Sakura to make songs. On February 12, 2002, as a place for exchanging users of Sakura, let's show off songs made with Sakura! For the purpose of being, a song bulletin board 3 was set up (* Older song bulletin board has been abolished since many works with problems in copyright were posted.).

1) From the toolbar above the editor, select the icon shown in Fig. 3.

2) Download the latest list. Since songs are frequently posted on the song bulletin board, select "Download latest list from WEB" from the "Download" menu as shown in Fig. 4.

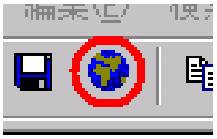
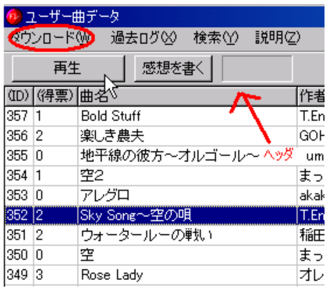


Fig. 3. Selected Icon in the Toolbar.



(ID)	(得票)	曲名	作者
357	1	Bold Stuff	T.En
356	2	楽しき農夫	GOH
355	0	地平線の彼方～オルゴール～	um
354	1	空2	まっ
353	0	アレグロ	akak
352	2	Sky Song～空の唄	T.En
351	2	ウォータールーの戦い	稲田
350	0	空	まっ
349	3	Rose Lady	オレ

Fig. 4. Download Menu in the Toolbar.

3) To listen to the song, move the cursor to the song title and press the [Play] button. At this time, by clicking on "header" at the top of the list, it is possible to rearrange the data in order of vote, song title, and author. In the song message board, if the number of songs posted exceeds 200 songs, the first 100 songs will move to the "past log". There are many wonderful songs among the songs that have moved to the past log. Therefore, if you select [Obtain past log list from WEB] from the past log menu, a list of past logs is displayed. If you select the past log from among them, it is possible to listen to the past log songs.

This software (Sakura) has a function to check this songboard bulletin board. It is nice feeling that the songs you make are adopted in games, etc. It is natural that people who cannot make music wish to add background music to the game. In this bulletin board, "I cannot make music but I want to add music as a homepage or BGM of the game!", "I use Sakura made songs as BGM, as a game, a website, etc. as BGM A MIDI material bulletin board as a place to interact with each other is desired.

MIDI material bulletin board MIDI data can be easily registered to the MIDI material bulletin board.

Quick help Double-click a word on the editor to display help.

In the editor of Sakura, when you double-click a word (or press the F1 key) with a mouse, the status bar shows how to use the command. For example, if you write Track on the editor, double-click the word with the mouse, or press F1 key, the word part is selected and quick help is displayed in the status bar. When help is displayed, if you want to see more detailed explanation, please click the status bar. Then, detailed help is displayed. Also, when you click on a variable, etc. the number of lines in which the variable is defined is displayed. At this time, clicking the status bar causes the cursor to jump to the variable definition line.

Control with DDE is possible to operate the Sakura editor from plugins and so on.

1) Use DDE in editor: If a language of which it is possible to use Dinamic Data Exchange (DDE) [15] that Windows uses for communication of applications. It is also possible to use Sakura.exe, a cherry editor, as a DDE reception (server).

[14] https://ja.wikipedia.org/wiki/Delphi

[15] https://en.wikipedia.org/wiki/Dinamic_data_exchange

2) Benefits: By operating the editor of Sakura from the program, it is possible to easily perform fixed form processing such as inserting and replacing sentences.

3) How to use: Here, we introduce an example of editor operation using my Japanese program language "Sunflower". To send a command to the Sakura Editor on a sunflower, use the instruction "send DDE". For example, to save the program you are currently editing, send "DDE" to "Sakura", "command", "save". I will write it.

(Rewrite the inside of the third "" to the following command.)

new: Initialize the editor
open File name: Opens the file specified in the file name.
save: Save the file currently being edited.
saveas file name: Save the program being edited to the file specified in the file name.
Insert string: Inserts a string at the current cursor position.
row Line number: Move the cursor to the line number.
col Column number: Move the cursor to the column number.
copy: Copies the currently selected text to the clipboard.
paste: Paste the contents of the clipboard into the editor.
seltext_save file name: Save the selected text to the specified file.
seltext_open file name: Rewrite the selected text to the contents of the specified file.
play: Play the contents of the editor.
stop: Stops playing.
Sample: "Replace selected text"
The following sample is a program that converts the part selected by the editor to capital letters. Selecting editor Transform text to uppercase:
DDE-transmit "copy" to "command" of "Sakura". '*** 1
Choose "Make text uppercase". If it is not, it will be ended.
Open clipboard. It converts it to uppercase. '*** 2
Save it to the clipboard. DDE sends "paste" to "command" of "himapad". '*** 3
end
*** 1: Instruct the cherry editor to copy the selection area to the clipboard.
*** 2: Convert the contents of the clipboard to uppercase
*** 3: Instruct the editor of Sakura to paste the contents of the clipboard to the selection area.

IV. Experiment

A. Conversion from Meteorological Data to Sound Data

Atmospheric pressure, wind direction, wind speed at the different altitude can be converted to numerical data as shown in Fig. 5. This is an example of text data of wind speed as function of time and atmospheric pressure. This is same things for the other meteorological data, wind direction, air temperature, relative humidity, etc. Thus, all kinds of meteorological data are represented as numerical text data. Then, these text data are converted to sound data of musical scale (melody), harmony, loudness, rhythm music instrument types as shown in Fig. 6. This is just an example for wind speed at the different altitude, or designated atmospheric pressure.

These are recorded on the track 1-4 depending on the altitude with the different music instrument types. On the other hand, wind direction is represented with musical scale (melody) depending on the eight different wind directions.

Sakura makes the following different types of sound features:

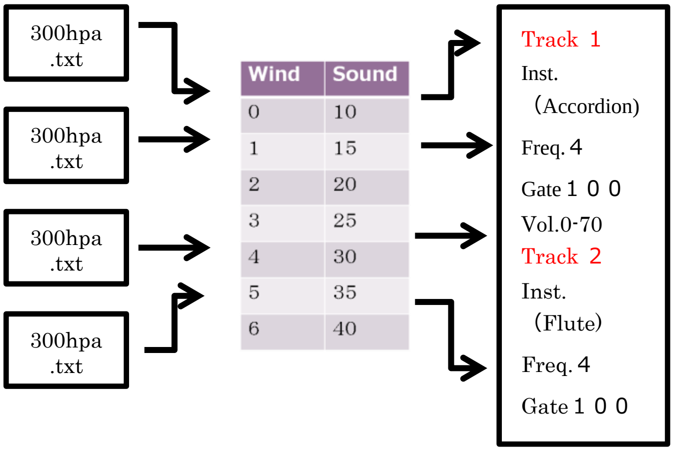
Rhythm, musical instrument type, melody, loudness,

Therefore, atmospheric pressure, air temperature, relative humidity, wind direction can be represented by input the text which is corresponding to the aforementioned meteorological data. Four different types of text input to Sakura are shown in Fig. 7. Thus the proposed system based on Sakura would help meteorologists for comprehensive understanding of meteorological and weather data and information in particular for weather forecasting.

This is just an example. There are so many other data and information which has to be represented by at least five data and information simultaneously.

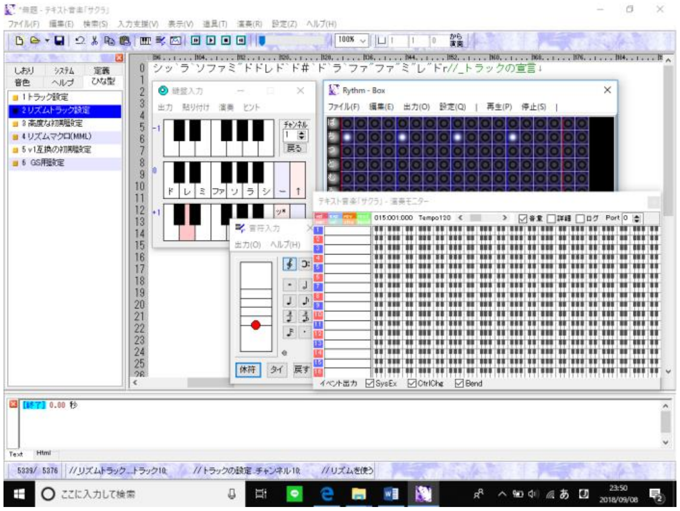
300hPa		500hPa		700hPa		850hPa	
Time	Wind	Time	Wind	Time	Wind	Time	Wind
6:00	1	6:00	1	6:00	1	6:00	1
0		5		0		8	
7:00	1	7:00	1	7:00	1	7:00	1
2		7		4		6	
8:00	1	8:00	1	8:00	1	8:00	1
3		8		5		5	
9:00	1	9:00	2	9:00	1	9:00	1
5		0		8		5	
10:00	1	10:00	2	10:00	1	10:00	1
7		2		9		7	

Fig. 5. Numerical Data of Atmospheric Pressure, Wind Direction, and Wind Speed.



Wind	Sound
0	10
1	15
2	20
3	25
4	30
5	35
6	40

Fig. 6. Conversion from Text File to Sound Data



```
シッ`ラ`ソファミ"ドドレド`ド#`ド`ラ`ファ"ファ"ミ"レ"ドr//_トラックの宣言↓
```

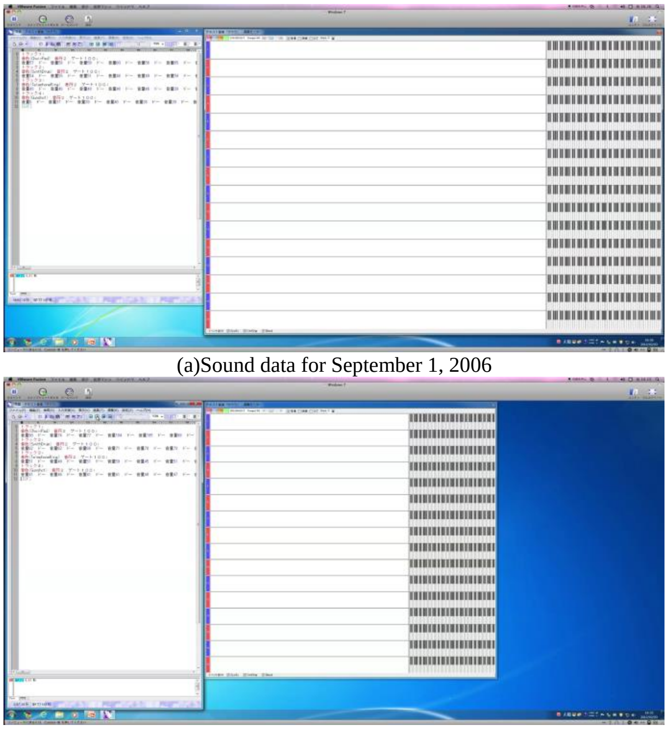
Fig. 7. Four Different Types of Meteorological Data Input to Sakura

B. Examples of Meteorological Satellite Data

Fig. 8(a) shows the edited sound data referring to text data which are converted from the TOVS/HIRS and MSU: Microwave Sounding Unit[16] derived meteorological data for September 11 to 22 in 2006 data while Fig. 8(b) shows those for December 1 to 12 in 2011. Although it is difficult to represent the sounds for both examples of meteorological data with this paper, it is confirmed that calm sound for December data while busy sound for September data. Also, it is confirmed the sounds of 3D meteorological data of air temperature, atmospheric pressure, relative humidity, wind direction and wind speed with the different music instrument types.

Values of meteorological data can be represented as volume (loudness) while the altitude can be represented as musical scale. The inverse relation between altitude and air temperature, relative humidity, atmospheric pressure can be identified with the sound.

Thus, user can hear the air temperature profile, relative humidity profile, atmospheric pressure of the pixel location in the MTSAT: Multi-functional Transport Satellite[17] image, for instance, displayed onto touch screen display by touching the pixel in concern. Most of meteorologist can understand the meteorological situation by location by location, it is hard to see the situation using just visual perception of the vertical profile images displayed onto screen though.



(a)Sound data for September 1, 2006

(b)Sound data for December 1, 2011

Fig. 8. Edited Sound Data Referring to Text Data which are Converted from the TOVS[18]/HIRS[19] and MSU[20] Derived Meteorological Data.

V. Conclusion

Sound User Interface: SUI with touch panel for representation of quantitative data and information together with its application to meteorological data representation is proposed. The proposed SUI is not a merely ear-con. Through experiments, it is found that the proposed SUI combined with visual perception makes meteorologist to understand meteorological data intuitively and is much understandable than ever in a comprehensive manner. It is also useful to hear "images" in particular, for blind person.

It is confirmed that the proposed SUI allows representation of meteorological data, air temperature, atmospheric pressure, relative humidity, wind direction and wind speed.

Further research works are required to expand the application fields of the proposed combined representation of weather data derived from remote sensing satellites. Also, application of the proposed SUI method will be for the blind persons.

[16] https://ja.wikipedia.org/wiki/TIROS-N/NOAA

[17] https://ja.wikipedia.org/wiki/MTSAT

[18] https://www.ozonelayer.noaa.gov/action/tovs.htm

[19] https://ja.wikipedia.org/wiki/TIROS-N/NOAA

[20] https://www.weblio.jp/wkpja/content/TIROS-N/NOAA_TIROS-N/NOAA の概要

ACKNOWLEDGMENT

The author would like to thank Mr. Tsuyoshi Morita of Saga University for his effort to conduct the experiments.

REFERENCES

[1] Shinichi Sobue, Kohei Arai, Hayato Okumura, Aya Yamamoto, Hiroshi Araki, Tsuneo Matsunaga, Proposed sonification system for representation of satellite remote sensing data, Journal of Space Research Information Analysis Society of Japan, 2011

[2] Kohei Arai, Sonification method for representation of multi-dimensional meteorological data derived from Earth observation satellites, International Journal of Research and Review on Computer Science, 3, 2, 1538-1542, 2012.

[3] Kohei Arai, Method for audible representation of meteorological data derived from remote sensing satellites, Proeedings of the SAI IntelliSys Conefrence 2018, to be published in the Springer Book Series, 2018.

[4] Kramer, Gregory, ed. (1994). Auditory Display: Audible, Audification, and Auditory Interfaces. Santa Fe Institute Studies in the Sciences of Complexity. Proceedings Volume XVIII. Reading, MA: Addison-Wesley. ISBN 0201626039.

[5] Flowers, J. H. (2005), Brazil, Eoin, ed., "Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions", Proceedings of the 11th International Conference on Auditory Display (ICAD2005): 406–409, http://www.icad.org/Proceedings/2005/Flowers2005.pdf

[6] d'Albe, E. E. Fournier (May 1914), "On a Type-Reading Optophone", Proceedings of the Royal Society of London

[7] Pollack, I. and Ficks, L. (1954), "Information of elementary multidimensional auditory displays", Journal of the Acoustical Society of America

[8] Chambers, J. M. and Mathews, M. V. and Moore, F. R. (1974), "Auditory Data Inspection", Technical Memorandum 74-1214-20

[9] Frysinger, S. P. (2005), Brazil, Eoin, ed., "A brief history of auditory data representation to the 1980s", Proceedings of the 11th International Conference on Auditory Display (ICAD2005) (Department of Computer Science and Information Systems, University of Limerick): 410–413, http://www.icad.org/Proceedings/2005/Frysinger2005.pdf

[10] "Continuous auditory monitoring--how much information do we register?", British Journal of Anaesthesia 83 (5): 747–749, 1999, doi:10.1093/bja/83.5.747, http://bja.oxfordjournals.org/content/83/5/747.full.pdf

[11] Kramer, G. and Walker, B.N. (2005), "Sound science: Marking ten international conferences on auditory display", ACM Transactions on Applied Perception (TAP) 2 (4): 383–388, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.7945&rep=rep1&type=pdf

[12] Thomas Hermann, Andy Hunt, and Sandra Pauletto. Interacting with Audible Systems: Closing the Loop. Eighth International Conference on Information Visualisation (IV'04) : 879-884. Available: online. DOI= http://doi.ieeecomputersociety.org/10.1109/IV.2004.1320244.

[13] Thomas Hermann, and Andy Hunt. The Importance of Interaction in Audible. Proceedings of ICAD Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia, July 6–9, 2004. Available: online

[14] Sandra Pauletto and Andy Hunt. A Toolkit for Interactive Audible. Proceedings of ICAD Tenth Meeting of the International Conference on Auditory Display, Sydney, Australia, July 6–9, 2004. Available: online.

[15] Stephen Barrass. Developing the Practice and Theory of Stream-based Audible. Journal of Media Arts Culture, scan, Available: online

AUTHOR'S PROFILE

Kohei Arai: He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission “A” of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 37 books and published 570 journal papers. He received 30 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.ht.

Local Average of Nearest Neighbors: Univariate Time Series Imputation

Anibal Flores[1], Hugo Tito[2], Carlos Silva[3]
E.P. Ingeniería de Sistemas e Informática, Universidad Nacional de Moquegua, Moquegua, Peru

***Abstract*—The imputation of time series is one of the most important tasks in the homogenization process, the quality and precision of this process will directly influence the accuracy of the time series predictions. This paper proposes two simple algorithms, but quite powerful for univariate time series imputation process, which are based on the means of the nearest neighbors for the imputation of missing data. The first of them Local Average of Neighbors Neighbors (LANN) calculates the missing value from the average of the previous neighbor and the following neighbor to the missing value. The second Local Average of Neighbors Neighbors+ (LANN+), considers a threshold parameter, which allows to differentiate the calculation of the missing values according to the difference between the neighbors: for the differences less than or equal to the threshold the missing value is calculated through of LANN and for major differences the missing value is calculated from the average of the four closest neighbors, two previous and two subsequent to the missing value. Imputation results on different time series are promising.**

Keywords—Univariate time series; imputation; LANN; LANN+

I. Introduction

Time series data are used in a large variety of real-world applications, and they often encounter the missing value problem due to data transmisión errors, machine malfunction, or human errors [1]. While imputation in general is a well-known problem and widely covered by different tools, finding algorithms or techniques able to fill missing values in univariate time series is more complicated [2]. The reason for this lies in fact is that the most imputation algorithms rely on inter-attribute correlations, while univariate time series imputation instead needs to employ time dependencies.

For univariate time series, the techniques that can be applied range from univariate algorithms, univariate time series algorithms and multivariate algorithms on lagged data [3].

In time series, can be find different gap sizes for NA values, a quick classification could be: short-gaps from 1 to 2 consecutive NAs; medium-gaps from 3 to 10 consecutive NAs; and big-gaps more than 10 consecutive NAs. In this paper we focus only on short-gaps.

In meteorological time series we find the three types of gaps mentioned above, we could even add a new category very big-gaps, since in some time series, there are gaps of NAs that range between approximately 1 and 72 months. A 72-month NA gap was found in the Punta de Coles time series between 1960/01/01 and 1965/12/31 (1978 consecutive NAs).

In this paper, we propose two algorithms for short-gaps of NAs within the univariate time series algorithms category and these are based on local averages of numerical time series. The first Local Average of Nearest Neighbors (LANN) algorithm is based on the average of the two nearest neighbors to the missing value or NA, the previous neighbor and the neighbor after the missing data. The second Local Average of Neighbors Neighbors+ algorithm (LANN+) is based on the difference (d) between the previous value and the value close to the missing value, this difference is compared with a threshold parameter that allows determining the way in which the missing value is calculated. When the differences are less than or equal to the threshold value, the missing value is calculated with the LANN algorithm and when the difference is greater than the threshold value, the missing value is calculated from the 4 neighbors closest to the NA value, the two previous and the two next to the NA value or missing value.

The paper is structured as follows: In Section II, a brief review of the state of the art regarding the proposals in this work is shown; in Section III, the fundamental theoretical bases for the better understanding of the paper content are shown; in Section IV, the proposed algorithms are described; in Section V, the results with different sizes of time series are described and discussed, likewise, they are compared with similar works; in Section VI, the conclusions reached at the end of the study are described and finally in the last Section VII, the future work is shown, which can be done to improve the proposals.

II. Related Work

A review of the state of the art of imputation works in univariate time series has been carried out and the results are shown below.

Commonly-used methods for univariate time series are relatively simple and include the arithmetic mean, interpolation, and last observation carried forward (LOCF) [4].

Last Observed Carried Forward LOCF [5] is a technique for replacing each NA with the most recent non-NA prior to it. For each individual missing value are replaced by the last observed value of that variable. In this work, zoo R package was used to implement LOCF imputation.

Hot-deck [6] imputation dates back to the days when data sets were saved on punch cards, the hot-deck referring to the “hot” staple of cards (in opposite to the “cold” deck of cards from the previous period). Most of the time, hot-deck imputation refers to sequential hot-deck imputation, meaning that the data set is sorted and missing values are imputed sequentially running through the data set line (observation) by line (observation). In this work VIM R package was used to implement hot-deck imputation.

Missing Value Imputation by Weighted Moving Average [7], the mean in this implementation taken from an equal number of observations on either side of a central value. This means for an NA value at position i of a time series, the observations *i-1,i+1* and *i+1, i+2* (assuming a window size of *k=2*) are used to calculate the mean. We have three types of algorithms in this category: Simple Moving Average (SMA), Linear Weighted Moving Average (LWMA) and Exponential Weighted Moving Average (EWMA).

Simple Moving Average (SMA) [2], all observations in the window are equally weighted for calculating the mean. For gap sizes equal to 1, and the parameter k equal to 1, SMA produces the same results as LANN in other cases results are different.

Linear Weighted Moving Average (LWMA) [2], weights decrease in arithmetical progression. The observations directly next to a central value i, have weight 1/2, the observations one further away (i-2,i+2) have weight 1/3, the next (i-3,i+3) have weight 1/4.

Exponential Weighted Moving Average (EWMA) [2] [8], is an approach that imputes the missing values by calculating the exponentially weighted moving average (EWMA). Initially, the value of the moving average window is set; the mean thereafter is calculated from equal number of observations on either side of a central missing value [8]. The observations directly next to a central value i, have weight $(1/2)^1$, the observations one further away (i-2,i+2) have weight $(1/2)^2$, the next (i-3,i+3) have weight $(1/2)^3$,.

In this work, imputeTS R package is used to implement SMA, LWMA y EWMA imputations.

Kalman Smoothing [8] on the state space representation of an autoregressive integrated moving average (ARIMA) model, is usually a good approach for imputation of highly seasonal univariate data [9]. In this work, we use imputeTS R package to implement ARIMA Kalman imputation.

Datawig[1] is a Python library that learns Machine Learning models using Deep Neural Networks to impute missing values in a dataframe. This method works very well with categorical and non-numerical features, therefore, it was not considered in the comparisons made in this work.

In order to compare the accuracy of the imputation techniques proposed with multivariable imputation techniques, two well-known multiple imputation algorithms were experimented, such as MICE [10] (Multiple Imputation by Chained Equations) and KNN [11] [12] (K-Nearest Neighbor), results can be seen in Section V.

III. Theoretical Background

A. Time Series

A time series is a set of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Some examples of time series are daily temperatures, weekly sales, customers per day, number of monthly visits, etc.

Studying the past behavior of a series will help to identify patterns and make better forecasts. When plotted, many time series exhibit one or more of the following features:

- Trends
- Seasonal and nonseasonal cycles
- Pulses and steps
- Outliers

B. Missing Data

Depending on what causes missing data, the gaps will have a certain distribution. Understanding this distribution may be helpful in two ways [3]. First, it may be employed as background knowledge for selecting an appropriate imputation algorithm. Second, this knowledge may help to design a reasonable simulator that removes missing data from a test set; such a simulator will help to generate data where the true values are known. Hence, the quality of an imputation algorithm can be tested.

Missing data mechanisms can be divided into three categories: Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR). In practice, assigning data-gaps to a category can be blurry, because the underlying mechanisms are simply unknown [3]. While MAR and MNAR diagnosis needs manual analysis of the patterns in the data and application of domain knowledge, MCAR can be tested for with t-test [3]

C. Univariate Time Series

A univariate time series is a sequence of single observations $o_1,o_2,o_3,\ldots$, on at sucessive points $t_1,t_2,t_3,\ldots\ t_n$ in time. Although a univariate time series is usually considered as one column of observations, time is in fact an implicit variable [3].

D. Univariate Imputation Methods

Techniques capable of doing imputation for univariate time series can be roughly divided into three categories [3]:

- Univariate algorithms. These algorithms work with univariate inputs, but typically do not employ the time series characteristics of the dataset. Examples are: mean, mode, median, random simple.
- Univariate time series algorithms. These algorithms are also able to work with univariate inputs, but make use of the time series characteristics. Examples of simple algorithms of this category are locf (last observation carried forward), nocb (next observation carried backward), arithmetic smoothing and

[1] W. Badr, “6 different ways to compensate for missing values in a dataset (data imputation with examples),”, Towards Data Science, [Online]. Available: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779. [Accessed 2019/07/15]

linearinterpolation. The more advanced algorithms are based on structural time series models and can handle seasonality.

- Multivariate algorithms on lagged data. Usually, multivariate algorithms cannot be applied on univariate data. But since time is an implicit variable for time series, it is possible to add time information as covariates in order to make it possible to apply multivariate imputation algorithms. This process is all about making the time information available for multivariate algorithms. The usual way to do this is via lags and leads. Lags are variables that take the value of another variable in the previous time period, whereas leads take the value of another variable in the next time period.

IV. Proposed Algorithms

A. *Local Average of Nearest Neighbors (LANN)*

LANN is an imputation algorithm for a univariate time series, which is fundamentally based on the average of the two closest neighbors, this according to the analysis carried out in several meteorological time series, where, it was observed that the previous neighbor v_{i-1} and the next neighbor v_{i+1} usually have approximate values at a certain value v_i. Where in an imputation problem v_i would be the NA value or the value to be imputed.

Table I shows the difference or distance between a time series value and the other values. The time series corresponds to meteorological data of maximum daily temperatures of 15 days at a weather station in the Moquegua Region, Ilo province from 2016-01-01 to 2016-01-15.

As mentioned earlier, this algorithm provides the same results as the SMA algorithm [2] when SMA is configured with the parameter k = 1 and the sizes of the gaps are equal to 1. When the size of the gaps is greater than 1 the results are different.

Then, from Table I, calculating the average of the diagonal elements that are exactly below the main diagonal or above, we will find the average difference between an element of the series and its first neighbor. Similarly, the following diagonal will give us the average difference between an element of the series and its second neighbor and so on. Table II shows the average differences for the 15-day time series.

According to Table II for the time series analyzed, we find that the closest neighbors to some value are 1st, 3rd, 6th, 9th and 2nd.

Next, we will experiment by generating random NA values in the previous time series and calculate the NA values by applying the average of the nearest neighbors (previous and next) with LANN algorithm according equation (1).

$$NA= (v_{i-1} + v_{i+1})/2 \qquad (1)$$

Table III shows the randomly generated NAs and their respective calculation using equation (1) with a percentage of missing data of 40%, 26.67%, 13.33%. The algorithm in Table IV was used in such a way that we make sure that we do not generate missing data at the beginning and at the end of the time series, likewise, the algorithm does not insert more than two NAs as gaps.

The LANN algorithm implemented in Javascript Language can be seen in Table V.

B. *Local Average of Nearest Neighbors+ (LANN+)*

LANN+ is based on the LANN technique, but it conditionally considers the average of the 4 closest neighbors instead of just two as in the LANN case. This algorithm uses a threshold parameter, which the higher it is, the imputation results will be very similar to the LANN algorithm. This parameter must be set according to the nature of the time series. For a temperature time series, the most appropriate is probably 1.0, in the case of an air passenger time series, the most suitable is probably 110.

TABLE. I. Matrix of differences between the Elements of a Time Series

	23.4	**22.8**	**22.6**	**23.4**	**24.4**	**24**	**23.6**	**25.2**	**24.4**	**23.6**	**23.8**	**24.2**	**23.8**	**24.8**	**24.8**
23.4	**0**	0.6	0.8	0	1	0.6	0.2	1.8	1	0.2	0.4	0.8	0.4	1.4	1.4
22.8	0.6	**0**	0.2	0.6	1.6	1.2	0.8	2.4	1.6	0.8	1	1.4	1	2	2
22.6	0.8	0.2	**0**	0.8	1.8	1.4	1	2.6	1.8	1	1.2	1.6	1.2	2.2	2.2
23.4	0	0.6	0.8	**0**	1	0.6	0.2	1.8	1	0.2	0.4	0.8	0.4	1.4	1.4
24.4	1	1.6	1.8	1	**0**	0.4	0.8	0.8	0	0.8	0.6	0.2	0.6	0.4	0.4
24	0.6	1.2	1.4	0.6	0.4	**0**	0.4	1.2	0.4	0.4	0.2	0.2	0.2	0.8	0.8
23.6	0.2	0.8	1	0.2	0.8	0.4	**0**	1.6	0.8	0	0.2	0.6	0.2	1.2	1.2
25.2	1.8	2.4	2.6	1.8	0.8	1.2	1.6	**0**	0.8	1.6	1.4	1	1.4	0.4	0.4
24.4	1	1.6	1.8	1	0	0.4	0.8	0.8	**0**	0.8	0.6	0.2	0.6	0.4	0.4
23.6	0.2	0.8	1	0.2	0.8	0.4	0	1.6	0.8	**0**	0.2	0.6	0.2	1.2	1.2
23.8	0.4	1	1.2	0.4	0.6	0.2	0.2	1.4	0.6	0.2	**0**	0.4	0	1	1
24.2	0.8	1.4	1.6	0.8	0.2	0.2	0.6	1	0.2	0.6	0.4	**0**	0.4	0.6	0.6
23.8	0.4	1	1.2	0.4	0.6	0.2	0.2	1.4	0.6	0.2	0	0.4	**0**	1	1
24.8	1.4	2	2.2	1.4	0.4	0.8	1.2	0.4	0.4	1.2	1	0.6	1	**0**	**0**
24.8	1.4	2	2.2	1.4	0.4	0.8	1.2	0.4	0.4	1.2	1	0.6	1	**0**	**0**

TABLE. II. AVERAGE DIFFERENCE BETWEEN A GIVEN VALUE AND ITS NEIGHBORS

Neighbour	**Average Difference**	**Ranking**
1st	0.6143	1st
2nd	0.8462	5th
3rd	0.6500	2nd
4th	0.8545	7th
5th	0.9600	9th
6th	0.7111	3rd
7th	0.8500	6th
8th	0.9143	8th
9th	0.7333	4th
10th	0.9600	10th
11th	1.3500	11th
12th	1.5333	13th
13th	1.7000	14th
14th	1.4000	12th

TABLE. III. RMSE OF THE LANN ALGORITHM (15 DAYS)

Real	NAs (40%)	LANN	NAs (26.67%)	LANN	NAs (13.33%)	LANN
23.4	23.4		23.4		23.4	
22.8	NA	23.00	22.8		22.8	
22.6	22.6		22.6		22.6	
23.4	23.4		NA	23.50	23.4	
24.4	NA	23.50	24.4		NA	23.70
24	NA	23.55	NA	24.00	24	
23.6	23.6		23.6		23.6	
25.2	25.2		25.2		25.2	
24.4	NA	24.50	24.4		24.4	24.40
23.6	NA	24.15	NA	24.10	23.6	
23.8	23.8		23.8		23.8	
24.2	NA	23.8	24.2		24.2	
23.8	23.8		NA	24.50	23.8	
24.8	24.8		24.8		24.8	
24.8	24.8		24.8		24.8	
RMSE		0.5041	RMSE	0.4330	RMSE	0.4950

TABLE. IV. RANDOM INSERTION ALGORITHM OF MISSING VALUES

```
function insertNAs(tso,k)
{    nts=tso.length;
     nn=(Math.floor(nts/k)-1);
     inf=1;sup=k;pna=0;
     for(j=0;j<nn;j++)
     {    pna=Math.floor(Math.random() * (sup - inf + 1)) + inf;
          pos.push(pna);
          tso[pna]="NA";
          inf+=k;
          sup+=k;
     }
     return tso;
}
```

TABLE. V. LANN ALGORITHM

```
function lann(tsna)
{    npos=pos.length;
     for(i=0;i<npos;i++)
     {    if(tsna[pos[i]-1]!='NA')
               prior=parseFloat(tsna[pos[i]-1]);
          else
               prior=parseFloat(tsna[pos[i]-2]);
          if(tsna2[pos[i]+1]!='NA')
               next=parseFloat(tsna[pos[i]+1]);
          else
               next=parseFloat(tsna[pos[i]+2]);
          base=(prior+next)/2;
          tsna[pos[i]]=base.toFixed(2);
     }
     return tsna;
}
```

The consideration of having a threshold is based on the fact that missing values in time series should not be imputed with the same technique, since each missing value and its neighbors have their own characteristics so there should be a technique that suits these characteristics in such a way that the imputed value has these characteristics. In that sense, although there is no exhaustive extraction of characteristics of time series with missing data, with LANN+, an effort is made to consider at least one characteristic that becomes the difference (d) between the previous neighbor and the neighbor after the missing value or NA data.

Regarding the alternation between two neighbors for differences less than or equal to the value of the threshold, and four neighbors for differences greater than the value of the threshold, it was considered so because when analyzing different time series of temperatures it was found that for small differences the average of the two closest neighbors (v_{i-1}, v_{i+1}) in most cases produced good results, while for larger differences it was more appropriate to use the average of the four nearest neighbors (v_{i-2}, v_{i-1}, v_{i+1}, v_{i+2}), something that can be seen if we compare the RMSE of Table III with those of Table VI.

TABLE. VI. RMSE OF THE LANN+ ALGORITHM (15 DAYS)

Real	NAs (40%)	LANN+ *	NAs (26.67%)	LANN+ *	NAs (13.33%)	LANN+ *
23.4	23.4		23.4		23.4	
22.8	NA	23.00	22.8		22.8	
22.6	22.6		22.6		22.6	
23.4	23.4		NA	23.35	23.4	
24.4	NA	23.50	24.4		NA	23.70
24	NA	23.55	NA	24.00	24	
23.6	23.6		23.6		23.6	
25.2	25.2		25.2		25.2	
24.4	NA	24.10	24.4		NA	24.05
23.6	NA	23.95	NA	24.10	23.6	
23.8	23.8		23.8		23.8	
24.2	NA	23.8	24.2		24.2	
23.8	23.8		NA	24.50	23.8	
24.8	24.8		24.8		24.8	
24.8	24.8		24.8		24.8	
RMSE		0.4873	RMSE	0.4308	RMSE	0.5534

*threshold=1.0

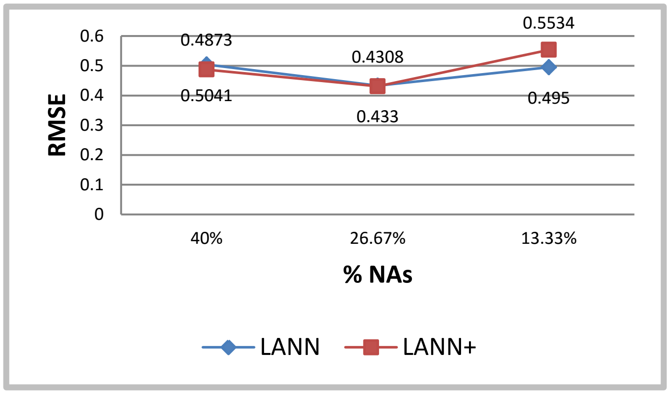


Fig. 1. RMSE Comparison between LANN and LANN+.

Fig. 1 shows a comparison between LANN and LANN+ for a time series of 15 days.

Table VII, shows the LANN+ algorithm implemented in Javascript language.

TABLE. VII. LANN+ ALGORITHM

```
function lannp(tsna)
{   npos=pos.length;
    for(i=0;i<npos;i++)
    {   if(tsna[pos[i]-1]!='NA')
            prior=parseFloat(tsna[pos[i]-1]);
        else
            prior=parseFloat(tsna[pos[i]-2]);
        if(tsna[pos[i]+1]!='NA')
            next=parseFloat(tsna[pos[i]+1]);
        else
            next=parseFloat(tsna[pos[i]+2]);
        d=Math.abs(prior-next);
        base=(prior+next)/2;
        if(d<=threshold)
            tsna[pos[i]]=base.toFixed(2);
        else
        {   mean2nn=get2nn_mean(tsna,pos[i]);
            tsna[pos[i]]=mean2nn.toFixed(2);
        }
    }
    return tsna;
}
```

V. RESULTS AND DISCUSSION

This section shows the results of comparing the proposed algorithms with other algorithms mentioned in section II. Likewise, the precision is compared with other time series with different characteristics to the temperature time series seen in Section IV.

A. Comparison with other Univariate Imputation Techniques

The LANN and LANN+ algorithms are compared with other imputation techniques in a maximum temperature time series of 15 days, Table VIII shows the results.

According to Table VIII, it is appreciated that for the percentage of NAs equal to 40%, the algorithm that obtained the best precision was the LWMA (0.4572) followed by the EWMA algorithm (0.4692) and thirdly the proposed algorithm LANN+ (0.4873). For the percentage of NAs equal to 26.67%, the algorithm with the best performance was the proposed algorithm LANN+ (0.4308) followed by LANN, SMA and ARIMA Kalman with the same RMSE (0.4330). For a percentage of NAs equal to 13.33%, in the first place, we have matched the LANN, SMA and ARIMA-Kalman algorithms with the same RMSE (0.4950).

Also, the performance of the same algorithms was evaluated with a time series with more data, in this case instead of 15 days, it is considered 90 days of maximum daily temperatures, from 2016-01-01 to 2016-03-30. Table IX shows the results.

According to Table IX, it can be seen that for a percentage of NAs of 48.89%, the algorithm with better precision is LANN (0.6059), secondly, we have the LANN+ algorithm (0.6196) and thirdly the SMA algorithm (0.6211). For a percentage of NAs of 32.22%, again the best precision was obtained by the LANN algorithm (0.5099), followed by the LANN+ algorithm (0.5296) and thirdly by the SMA algorithm (0.5451). For a percentage of NAs of 23.33%, the best algorithm was EWMA (0.4765), followed by LWMA (0.4970) and thirdly we have two, LANN and SMA with a RMSE equal to 0.5085.

The proposed algorithms were also compared with the precision of two well-known multiple imputation algorithms such as MICE and KNN and the results shown in Table X were obtained. In this case, it's used the data from the same previous data range of the nearest meteorological station to the Punta de Coles station, which is the Ilo Station. Ilo station is located in the El Algarrobal district of the province of Ilo.

TABLE. VIII. COMPARISON WITH OTHER UNIVARIATE IMPUTATION TECHNIQUES – 15 DAYS

Technique	RMSE (NAs 40%)	RMSE (NAs 26.67%)	RMSE (NAs 13.33%)
LANN	0.5041	**0.4330**	**0.4950**
LANN+ (treshold=1.0)	**0.4873**	0.**4308**	0.5534
LOCF	0.8869	0.6324	0.9055
Hotdeck	0.9201	1.0295	0.8000
SMA	0.6448	**0.4330**	**0.4950**
LWMA	**0.4572**	0.4721	0.6275
EWMA	**0.4692**	0.4613	0.6170
ARIMA Kalman	0.5482	**0.4330**	**0.4950**

TABLE. IX. COMPARISON WITH OTHER UNIVARIATE IMPUTATION TECHNIQUES – 90 DAYS

Technique	RMSE (NAs 48.89%)	RMSE (NAs 32.22%)	RMSE (NAs 23.33%)
LANN	**0.6059**	**0.5099**	**0.5085**
LANN+ (1.0)*	**0.6196**	**0.5296**	0.5210
LOCF	0.8382	0.7485	0.7461
Hotdeck	1.322	1.5349	0.5778
SMA (k=1)	**0.6211**	**0.5451**	**0.5085**
LWMA (k=4)	0.6428	0.6299	**0.4970**
EWMA (k=4)	0.6266	0.5878	**0.4765**
ARIMA Kalman	0.7447	0.5964	0.5191

* threshold=1.0

TABLE. X. COMPARING WITH MICE AND KNN

Technique	RMSE (NAs 48.89%)	RMSE (NAs 32.22%)	RMSE (NAs 23.33%)
LANN	0.6059	0.5099	0.5084
LANN+*	0.6196	0.5296	0.5210
MICE	1.4208	1.0993	0.8632
KNN	1.0842	0.9762	0.9646

*Threshold: 1.0

According to Table X, the accuracy of the proposed LANN and LANN+ algorithms greatly outperform MICE and KNN.

The proposed algorithms were also evaluated with time series with other characteristics:

- Airpass: Monthly total international airline passengers from 01/1960-12/1971 [13] Characteristics: trend, seasonality.
- Beersales: Monthly beer sales in millions of barrels, 01/1975-12/1990 [13] Characteristics: no trend, seasonality.

Table XI shows the results achieved with the Airpass time series.

Table XI shows that for time series with different characteristics than maximum temperatures, the proposed algorithms also offered good performance.

Table XII shows the results with the Beersales time series, where the LANN algorithm showed the best accuracy in the imputation process of missing data.

TABLE. XI. COMPARISON ON AIRPASS TIME SERIES

Technique	**RMSE**
LANN	**22.0368**
LANN+*	**20.9122**
LOCF	43.6041
Hotdeck	164.6075
SMA	**21.7995**
LWMA	28.9395
EWMA	24.4703
Kalman-ARIMA	**20.8952**

*Threshold: 110

TABLE. XII. COMPARISON ON BEERSALES TIME SERIES

Technique	**RMSE**
LANN	**0.8738**
LANN+ *	0.9738
LOCF	1.6869
Hotdeck	2.6295
SMA	**0.9246**
LWMA	1.1915
EWMA	1.0772
Kalman-ARIMA	**0.9283**

*Threshold: 0.02

VI. CONCLUSIONS

The proposed algorithms showed a very good performance in the imputation process of NAs short-gaps in different time series in which they were analyzed. They outperformed many well-known imputation algorithms such as ARIMA-Kalman, Hotdeck, LOCF, MICE, KNN in different percentages of missing data.

For meteorological time series such as maximum temperature series, LANN and LANN+ are highly recommended due to the good accuracy achieved.

For the time series with high trend and seasonality, the use of the LANN+ algorithm is recommended and for time series with low trend and high seasonality, the use of LANN is recommended.

VII. FUTURE WORK

The algorithms proposed in the present work have been analysed and evaluated in short-gaps of NAs, it is important in future works to configure them for larger gaps, three or more data and evaluate the corresponding accuracy.

The proposed algorithms can be improved by combining with forecast models such as Deep Learning, especially Recurrent Neural Networks [14] especially Long Short Term Memory (LSTM) or Gate Recurrent Unit (GRU) that allow improving the accuracy of the estimates reached.

REFERENCES

[1] Chang, C. Wang, S. Lee, "Novel Imputation for Time Series Data," in International Conference on Machine Learning and Cybernetics, Guangzhou, 2015.

[2] S. Moritz, T. Bartz-Beielstein, "imputeTS: Time Series Missing Value Imputation in R," The R Journal, vol. 9, no. 1, pp. 207-2018, 2017.

[3] S. Moritz, A. Sardá, T. Bartz-Beielstein, M. Zaefferer, J. Stork, "Comparison of different Methods for Univariate Time Series Imputation in R," arxiv.org, 2015.

[4] N. Bokde, M. Beck, F. Martinez, K. Kulat, "A novel imputation methodology for time series based on pattern sequence forecasting," Pattern Recognition Letters, 2018.

[5] A. Zeileis, G. Grothendieck, "zoo: S3 infrastructure for regular and irregular time series,"Journal of Statistical Software, vol.14, no. 6, 2005.

[6] A. Kowarick, M. Templ, "Imputation with the R package VIM," Journa of Statistical Software, vol. 74, no. 7, 2016.

[7] S. Moritz, "Package ImputeTS," cran.r-project.org, 2019.

[8] E. Rantou, "Missing Data in Time Series and Imputation Methods," University of the Aegean, Samos, 2017.

[9] A. Chaudhry, W. Li, A. Basri, F. Patenaude, "On improving imputation accuracy of LTE spectrum measurements data," in Wireless Telecommunications Symposium, Phoenix, AZ, USA, 2018.

[10] S. Van Buuren, K. Groothuis-Oudshoorn, "mice: multivariate imputation by chained equations in R," Journal of Statistical Software, vol. 45, no. 3, 2011.

[11] G. Chang, T. Ge, "Comparison of missing data imputation methods for traffic flow," in International Conference of Transportation, Mechanical, and Electrical Engineering (TMEE), Chanchung, China, 2011.

[12] B. Sun, L. Ma, W. Cheng, "An improved k-nearest neighbours method for traffic time series imputation," in Chinese Automation Congress (CAC), 2017.

[13] K. Chan, B. Ripley, "TSA: Time series analysis," CRAN. R-project.org, 2012.

[14] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values," Scientific Reports, 2018.

Mortality Prediction based on Imbalanced New Born and Perinatal Period Data

Wafa M. AlShwaish[1], Maali Ibr. Alabdulhafith[2]
College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

***Abstract*—This study was carried out by the New York State Department of Health, between 2012 and 2016. This experiment relates to six supervised machine learning methods: Support Vector Machine (SVM), Logistic Regression (LR), Gradient Boosting (GB), Random Forest (RF), Deep Learning (DL) and the Ensemble Model, all of which are used in the prediction of infant mortality. This experiment applied ensemble model that concentrated on assigning different weights to different models per output class in order to obtain a better predictive performance for infant mortality. Efforts were made to measure the performance and compare the classifier accuracy of each model. Several criteria, including the area under ROC curve, were considered when comparing the ensemble model (GB, RF and DL) with the other five models (SVM, LR, DL, GB and RF). In terms of these different criteria, the ensemble model outperformed the others in predicting survival rates among infant patients given a balanced data set (the areas under the ROC curve for minor, moderate, major and extreme were 98%, 95%, 92% and 97% respectively, giving a total accuracy of 80.65%). For the imbalanced dataset, (the areas under the ROC curve for minor, moderate, major and extreme were 98%, 98%, 99% and 99% respectively, giving total accuracy increased to 97.44%). The results of the experiments used in this dissertation showed that using the ensemble model provided a better level of prediction for infant mortality than the other five models, based on the relative prediction accuracy for each model for each output class. Therefore, the ensemble model provides and extremely promises classifier in terms of predicting infant mortality.**

***Keywords*—Component; machine learning; support vector machine; logistic regression; gradient boosting; random forest; deep learning; ensemble model**

I. INTRODUCTION

In high-income countries, a significant part of public spending is committed to prevention and care. Chronic illnesses such as cancer, asthma and high fevers present serious barriers to infant survival, and dramatically increase spending on healthcare services. A World Health Organization survey in 2017 estimated that children under five years accounted for 5.4 million of these deaths. The inpatient discharge information has been play a crucial role in improving understanding of risk factors mechanisms in infants. Recent approaches using data mining techniques and machine learning algorithms have become part of the experimental processes of many disciplines over recent years [1].

In the healthcare field, data mining techniques help researchers analyze very large databases, and are useful in directing hospital policies to increase patient flow and minimize non-value-added care time. Classifications based on statistical analysis and Artificial Neural Networks, as well as other Machine Learning algorithms are becoming more common aspects of predictive healthcare models [3]. To increase the accuracy in prediction of machine learning models, new variables relating to patient information are used to construct the models. This study applies machine learning in an effort to change the current healthcare process from a receptive model into one that is increasingly proactive. The research question to be answered in this study can be stated as:

RQ1: Can we identify the features that help to predict infant mortality?

However, no study has yet applied ensemble machine learning methods that assign different weights to different models per output class.

II. REVIEW OF EXISTING LITERATURE

A. Infant Mortality

Identifying the variables which affect statistics on health can be used to predict and thereby address and improve long-term survival rates for infants. Kong et al. (2016) identified various predictors of mortality and morbidity among infants, pinpointing many factors relating to pre-term births aside from gestation and birth weight that could be associated with risks relating to high mortality and morbidity risks. For example, infants born at 24 and 25 weeks were more likely to die than infants born more than 26 weeks into the pregnancy. Identifying infants' risk levels by predicting future health outcomes helps improve the efficiency and quality of health care [2].

Diagnoses relating to the level of infant risk are important in terms of both clinical decision making and the provision of care for newborn children a study by Martinez (2017) identified predictors for prolonged hospitalization or readmission for acute lower respiratory infections (ALRIs) in infants with bronchopulmonary dysplasia (BPD). This wide-ranging study was conducted using nationally representative data from children on a US inpatient database, and included a total of 138 patients. The study used logistic regression with and without an interaction term between gender and breastfeeding. The results of the regression showed a p value of ≤ 0.05 and odds ratios (OR) at 95% confidence intervals [3]. Studies have also proved that lower neonatal mortality rates were associated with early breastfeeding compared with higher mortality rates for late breastfeeding [4].

B. *Machine Learning*

The following will provide an overview of the various methods that are widely classified as supervised and unsupervised machine learning techniques in predicting infant mortality risks.

Unsupervised machine learning in used to collect data with similar attributes into groups. Sample testing is then classified based on proximity within these groups. The groups are generated based on similarity scales such as probabilistic or Euclidean distance. Ravishankar and Clarke (2017) described commonly used clustering techniques and applied the iterative k-means clustering algorithm, in which the outliers in the legend are used to efficiently identify clusters on Dept. of Health of New York State. The algorithms proved successful in terms of processing for data analysis framework by applying data cleansing/ETL, data joining, classification and prediction, visualization of results, interpretation and reporting. Outliers in cost increases (such as the Monroe Community Hospital) were identified through iterative k-means clustering [5].

Supervised classification techniques are the most commonly implemented methods employed by intelligent systems, and are applicable to cases which contain labeled data. Kenley and Shimony (2016), provide insights into predicting the brain maturity of infants and adolescents using structural and functional magnetic resonance imaging data. Data were evaluated throughout the duration of the functional magnetic resonance imaging of 50 new births from Louis Children’s Hospital Neonatal Intensive Care Unit (NICU). The results of the experiment showed that using the Support Vector Machine (SVM) model to achieve results improved accuracy, sensitivity, specificity and p-values for binomial probabilities [6]. Previous studies used data from NICU patient data systems. Rinta-koski and Simo (2017), explored powerful mortality classification models using the SVM and Gaussian Process (GP) to identify clinical features on arrival at the Neonatal Intensive Care Unit and those made during the first 72 hours of care for 598 Very Low Birth-Weight infants (birth-weights of below 1500g), with combined features extracted from sensor measurements. The SVM achieved better classification accuracy (0.931) [7].

III. Experimental Design and Methodology

This section sets out the nature of the experiments that will be used to answer the research question. The CRISP-DM methodology offers a structured approach to data mining [8]. The study will be performed as a five-stage process including evaluation, data understanding, data preparation, modelling and evaluation. Each step in the study will be undertaken using Python programming language and the Tensorflow library [9], an open source library for fast numerical computing, and the Scikit-learn library [10], an open source machine learning library for the Python programming language characterized by several classification, regression and clustering algorithms. The section is divided into sub-sections based on the CRISP-DM framework as shown in Fig. 1, each of which will cover the framework in more detail.

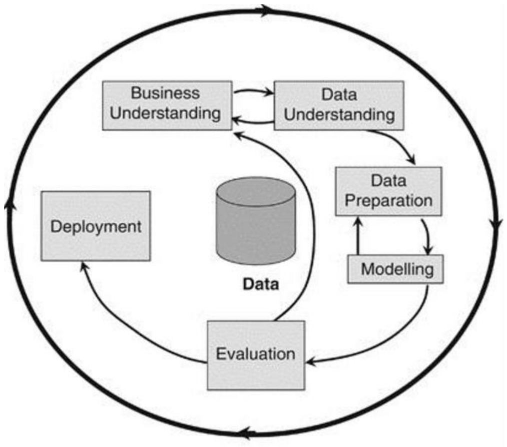


Fig. 1. CRISP-DM Model.

The high-level experiment is illustrated in below Fig. 2.

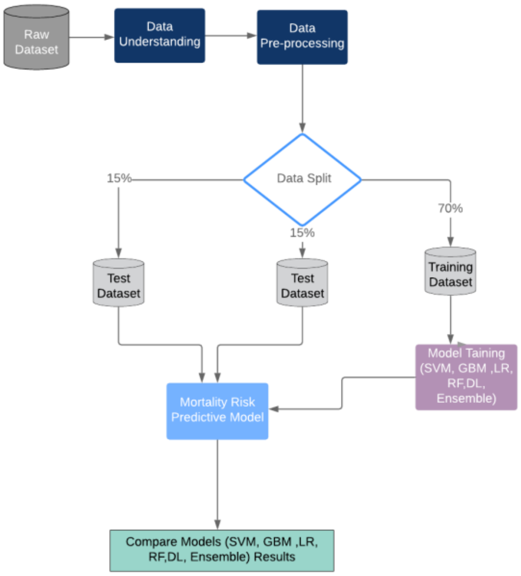

Raw Dataset
Data Understanding
Data Pre-processing
15%
Data Split
70%
15%
Test Dataset
Test Dataset
Training Dataset
Model Taining (SVM, GBM ,LR, RF,DL, Ensemble)
Mortality Risk Predictive Model
Compare Models (SVM, GBM ,LR, RF,DL, Ensemble) Results

Fig. 2. High Level Design Experiment.

IV. Implementation

This section describes the results of the study and the experiments that were performed. The section pattern emulates that of the Design and Implementation section to make comparisons easier between balanced and unbalanced dataset outcomes.

A. *Exploratory Analysis of Dataset*

This section explores the infant and perinatal period dataset. First, we need to eliminate the duplicate 'Live born' words from the CCS diagnosis description feature since all patient infants were born alive, and the word does not add much meaning. It can therefore be safely ignored. However, we explored this feature to check the effectiveness of it on each class of mortality. Fig. 3 shows that there are differences between diagnoses in each class of mortality, although the Minor class is similar to entire population because this class represents 97% of data set.

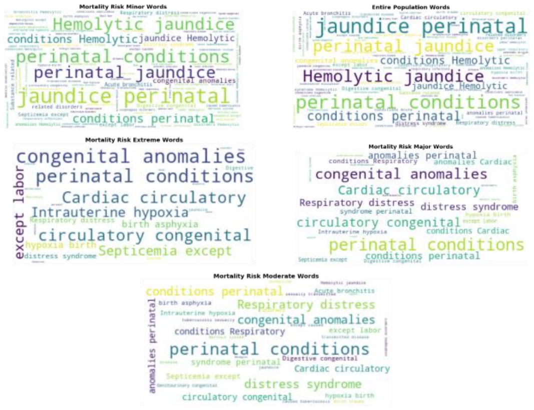


Fig. 3. Entire Population CCS Diagnosis Description Words.

As Fig. 4 shows, the mortality risk distribution by gender indicates that women enjoy better healthcare because the Extreme and Major figures are below the average figure for all patients. Meanwhile, males take less health care because their Extreme and Major cases are above the average patient care data.

However, Fig. 5 and 6 shows that white patients enjoy better healthcare than other races because extreme, moderate and major figures are less than the average in terms of patient care. In addition, the broad range of other ethnicities generally has worse than average health. This appears to indicate discrimination in healthcare depending on race and ethnicity in the USA [11].

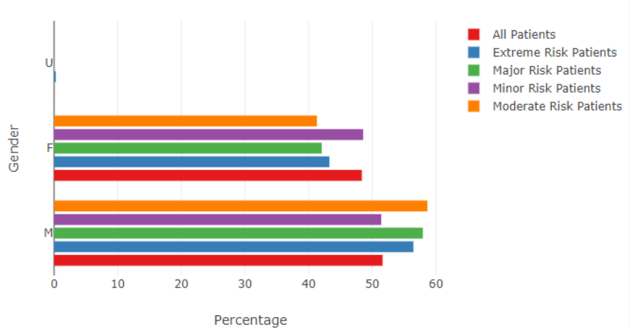


Fig. 4. Mortality Risk Distribution by Gender.

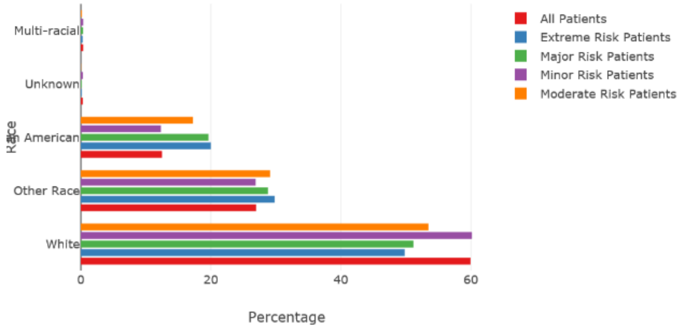


Fig. 5. Mortality Risk Distribution over Race.

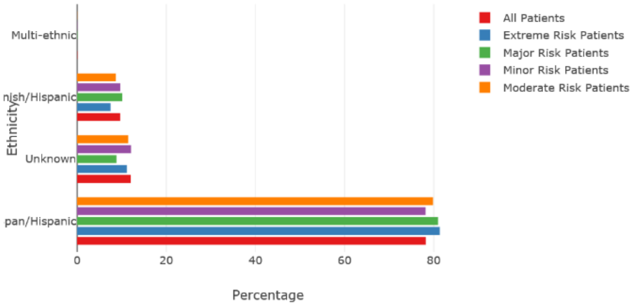


Fig. 6. Mortality Risk Distribution over Ethnicity.

B. Understanding the Data

Statistics were generated to analyse the data in each column to discover a number of different values for each column. From the analysis, it was immediately apparent that the Age Group variable is not important as there is only one value (0 to 17) in the dataset. We will therefore not use this column. In the dataset, the birth weight column has a zero value for some records where we need to analyse mortality rates for new born infants. We checked to see why one row includes zero values for the Birth Weight column. It appears to indicate that a zero birth weight entry means that the patient might not be a new born, but we need to investigate this more deeply to confirm the assumption. After checking the Diagnosis column which also includes records with a zero birth weight value, we observed different diagnoses. That mean records have had zero birth weights entered by mistake, so we need to remove these zero values. However, in order to analyse missing data, we found missing data in only two features (see Table I), so we distinguish those features by assigning a value of -1.

Before we proceed with the analysis, we need to compare different kinds of dimensionality reduction for plotting purposes, since we have 16 dimensions that we need to reduce to 2 dimensions in order to obtain good reduction plotting. Because most of the data is dense, we used principal component analysis (PCA) and t-SNE. Fig. 7 shows the PCA results, in which point distribution is not clear.

However, the t-SNE analysis shown in Fig. 8 shows Extreme cases as clear areas, thereby offering better results than PCA. We then took the process through different kinds of outlier detection algorithms to check outliers from the data such as Robust Covariance, One-class SVM and Isolation Forest. We began with the Robust Outlier detector shown in Fig. 9, in which all points outside the boundary ellipse are outliers.

TABLE. I. MISSING VARIABLE

Variables	Count
Operating provider license number	45681
Other provider license number	159058

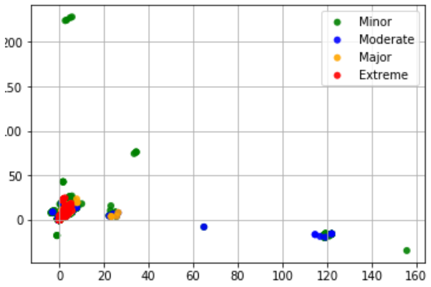


Fig. 7. PCA Data Distribution Plot.

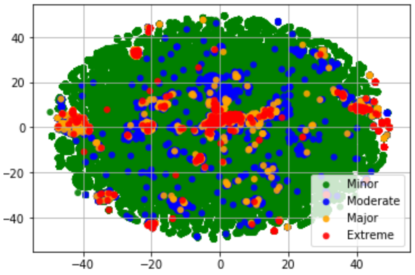


Fig. 8. t-SNE Data Distribution Plot.

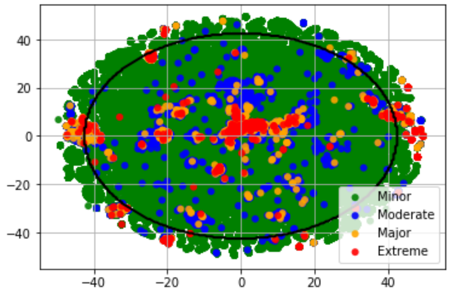


Fig. 9. t-SNE Data Distribution Plot with Robust Outlier Detector.

We then used the Isolation Forest algorithm shown in Fig. 10 which defines a separate boundary between points from outliers of all classes.

However, we needed to perform another kind of outlier detection test. This was the One-class SVM shown in Fig. 11 that sets regional boundaries; all points inside boundaries are valid data, while those outside it is outliers. The One-class SVM was the best detector of outliers, and data within the boundaries are consistent and closed.

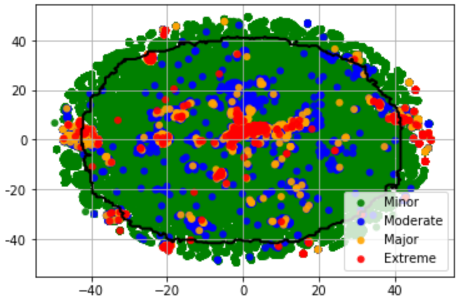


Fig. 10. t-SNE Data Distribution Plot with Isolation Forest Outlier Detector.

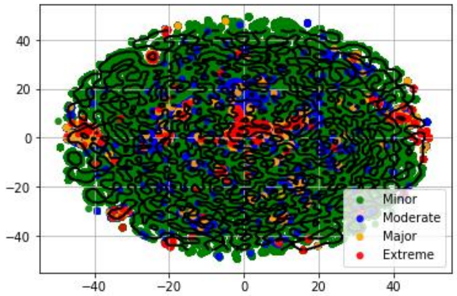


Fig. 11. t-SNE Data Distribution Plot with One-Class SVM Outlier Detector.

Cramer's V Association was then used to interpret associated factors between the nominal variables. The association range lies between 0 and 1, and greater values show stronger associations. The correlation matrix for Cramer's V Association heatmap matrix is shown in Fig. 12, which shows the relationship between the features. The result obtained from matrix is as follows:

- No single feature is strongly associated with APR risk of mortality.
- Patient Disposition and APR Severity of Illness Code have strong positive associations with APR risk of mortality.

The uncertainty coefficient was also used in the study to explain associations between categories. The correlation coefficient determines the degree of association between two variables, and this is shown in Fig. 13, which shows the relationship between features. As shown in both associations the APR Severity of Illness description has a strong positive association with the APR risk of Mortality.

C. Data Preparation

After the analysis of the data is complete, the next step is to remove those issues that have been identified in the dataset so that the remaining data will fit the processes used in modelling.

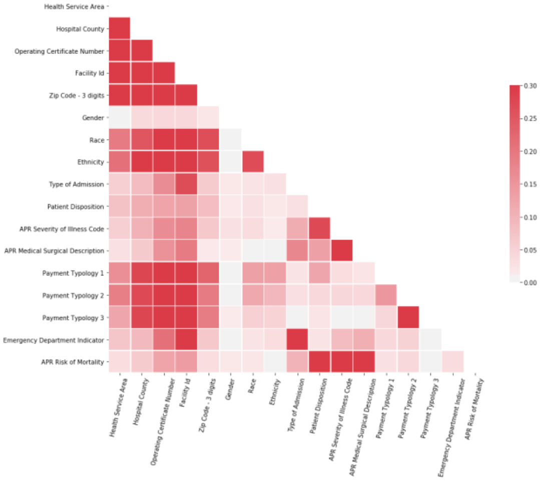


Fig. 12. Cramer's V Association Matrix of Variables with APR risk of Mortality.

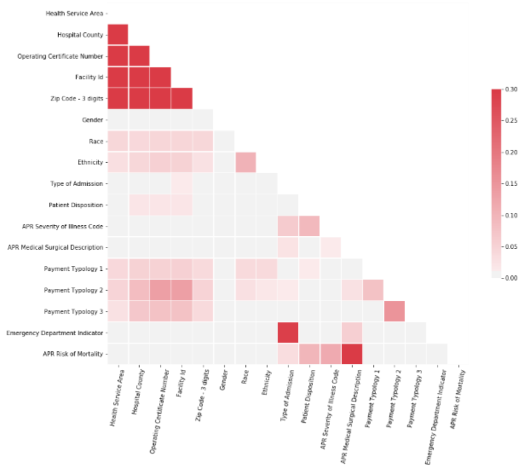


Fig. 13. Uncertainty Coefficient of Variables with APR Risk of Mortality.

1) Balanced dataset: Under Sample:Classifier machine learning algorithms such as Random Forest tend to give results biased towards classes which have the highest number of records.

Classifier algorithms can ignore the features of minimal class, considering them no more than noise. It is highly probable that minimal classes will be misclassified when compared to better populated ones. The reason for using 'Pandas sample' is because we have imbalance between mortality classes, as shown in Fig. 14. The 'Pandas sample' is implemented on the imbalanced dataset samples to balance it. The number of records showing extreme mortality risk is much lower than numbers in the other classes, as shown in Table II.

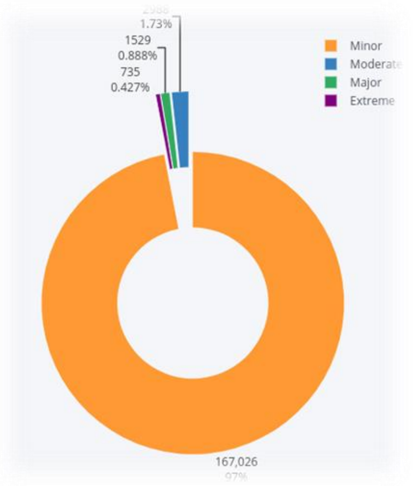


Fig. 14. Unbalanced Data Set.

TABLE. II. UNDER SAMPLING

Target: Mortality Risk	Imbalanced Data set	Balanced Dataset
Minor	167026	735
Moderate	2988	735
Major	1529	735
Extreme	735	735

2) Standardization standard scaler: The distribution of the Birth Weight feature as shown in Fig. 15 provides information about infants' weights. As shown in histogram, the average birth weight is 3.5 kilograms, and birthweights go up to 5 kilograms. Because of this, we need to test for normality using a variety of statistical analyses.

First, we used the Shapiro-Wilk test, which returned a p-value of 0.00, which is less than .05. We then used the Normal-t test, which also returned a value of 0.00. We also used the Anderson-Darling test to see if our data came from a normal distribution. The null hypostudy was rejected, similar to the previous two tests.

The QQ plot could provide us with more certainty about the normality, and also offers better visualization. From the QQ plot shown in Fig. 16 we can see how the data appears, and it is immediately apparent that the data are not normally distributed. This visualization helps us to study abnormal cases in our experiment.

3) Encoding categorical variables: After a balanced dataset had been successfully created, the only problem that remained was to remove categorical variables, as most machine learning models work only on numeric variables and cannot compute using features containing string values. All 18 of the categorical variables were nominal, and this meant that the values within those categorical variables did not follow a specific natural order. In order to remove nominal variables, we performed encoding procedure.

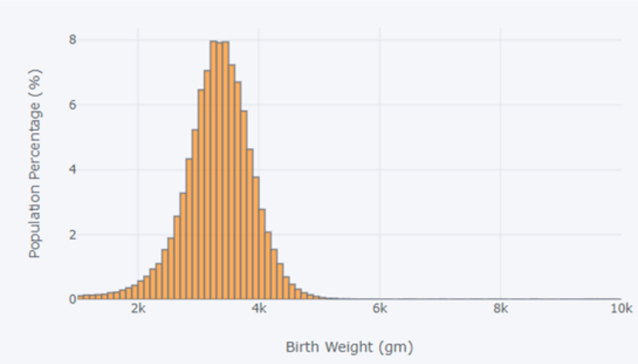


Fig. 15. Entire Population Birth Weight.

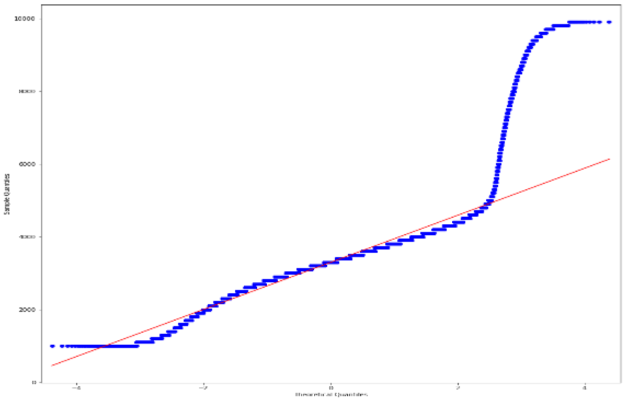


Fig. 16. QQ Plot for Birth Weight.

4) Text vectorization: This part of the analysis focused on creating data vectors from text vectorization by importing a TF-IDF Vectorizer from sklearn.feature_extraction.text. The vectorizer was initialized, fitted and transformed to calculate the TF-IDF score for the text in the [(CCS Diagnosis Description)] feature. The sklearn fit_transform performed both fit and transform functions, and the output took the form of a skewed matrix.

D. Models

In this phase, we built classification models to predict infant mortality risk using Gradient Booted (GB), Support Vector Modelling (SVM), Random Forest (RF), Logistic Regression (LR), Deep Learning (DL) and ensemble models. The balanced data set was created after adding and encoding categorical data and normalizing the results to use for the construction of models. Before training the model, input data regarding text vectorization and strength of association of features was implemented to obtain a better fit for the model. For each model we processed data in two ways: first, we used the scikit-learn train_test_split method to divide the dataset into training, validation and testing datasets that would maintain the distribution of the output. 70% of the data were used for training, 15% for validation and 15% for testing.

1) Logistic regression: Logistic Regression was used to provide a multi-class classification regression model. First, the LR module was imported to create an LR classifier object using the Logistic Regression cross validation function to get best parameters.

2) Gradient boosted tree: We used XGBoost to predict the mortality risk for infants. We imported XGBoost, which uses an assessment metric to check the performance of the Training Model on the test dataset.

3) Random forest: The Random Forest (RF) model used a randomized search function to evaluate the best hyper-parameters. While the parameters were learned during the model training, hyper parameters must be set before training. The importance of each feature in the RF classification is indicated by the sum of the reduction in Gini Impurity (a measure that the decision tree uses to minimize when splitting each node for every node that is split by that feature. We can use these to attempt to calculate which of the predictor variables the RF considers most important in terms of mortality risk. The feature importance can be extracted from a trained RF.

4) Support vector classifier model: We applied a prepackaged model provided by a scikit-learn support vector classifier to train an SVM model on this data. Tuning parameter values for machine learning algorithms effectively improves the performance of the model.

5) Deep learning: We used fully connected network architecture to implement our infant mortality risk prediction model. We used the Class Weight variable to calculate the class weight and added it to model. After the model had been created, we were able to make predictions according to all the learned nodes.

6) Ensemble model: Most of the known ensemble techniques do not account for the relative prediction accuracy of multiclass classification problems. It either uses a blanket weighting for all classes or uses voting, which could lead to equal votes for multiple different outputs, as shown in Fig. 17.

In our approach, we decided to feed into the deep neural network the output probability per class from the different models, as they would allow the deep neural network to give different weights to different models per output class. This improved the overall prediction accuracy, which was based on the relative prediction accuracy for each model per output class as shown in Fig. 18.

One limitation in most previous studies is that they only considered a blanket weighting for all classes, and no previous study has concentrated on assigning different weights to different models per output class.

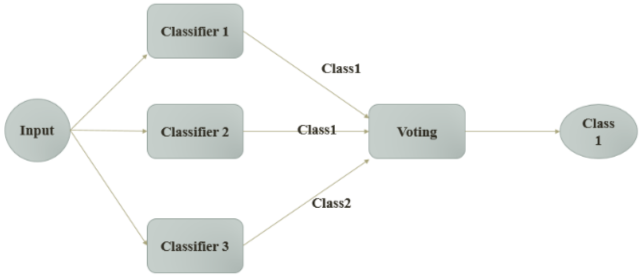


Fig. 17. Voting Mechanism.

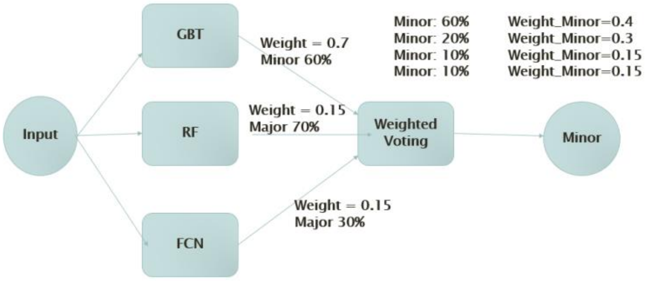


Fig. 18. Voting Mechanism with different Weights.

V. Evaluation and Results

A detailed analysis of the experiments described in the previous section will be provided in this section, including the results of each experiment. The experiments were performed in order to build six models for supervised machine learning. This section evaluated the execution of each model according to the levels of accuracy gained after running each experiment on the dataset. The same experiment was also performed on the imbalanced dataset which contained biased values relating to the mortality risk features. When evaluating the performance of models, box plots were created using confusion matrixes that summarized the prediction results generated as well as the accuracies achieved. We also used cross validation techniques by applying a series of training/validation/test set splits based on logistic regression and random forest methods. The statistical analysis of the study result will also be discussed in this section.

A. Comparison of Average Performance of Imbalanced and Balanced Target Data

Imbalanced data sets are common in predictive classification experiment. Fig. 19 shows the results for each classifiers on the imbalanced dataset in which the mortality risk was highly biased towards instances having a ‘minor’ classification, for all models.

The first analysis was performed on the results gained from the unbalanced data set. It is clear from the histogram of the accuracies that random forest and the ensemble model have higher accuracies than the models derived from the other algorithms. The maximum F1 score obtained by random forest model for the ‘extreme’ class is above 60%.

Due to the imbalanced data, the under-sampling approach has been taken to create a balanced dataset. Using alternative metrics like F1 score, recall, precision, true positive rate and false positive rate is strongly recommended in place of using the accuracy of the model to measure its performance.

Tables III and IV shows the mean accuracies and classification metrics obtained from the imbalanced dataset and balanced data set of each model.

B. Comparison of Classifiers Performance

A further experiment was performed on all six models after applying the under-sampling technique. Fig. 20 shows the performance histogram of balanced dataset models. The most remarkable change here is the increase in F1 score for each target variable value. As the graph shows, the random forest model and the ensemble model have higher prediction accuracy when compared to other models.

C. Strengths and Limitations of Results

Machine learning algorithms and their use was considered as an integral factor in the research. The experiment used six machine learning algorithms (LR, RF, GB, DL, Ensemble and SVM), which were similar in the ways in which they were used for the classification of variables.

The training models that were relevant to the different families in the same data set can be seen as one of the strengths of the study. Likewise, the ensemble model was built using different models–random forest, gradient boosting and deep learning–in order to obtain an efficient performance from the model. In the ensemble model, we decided to feed the output probability per class from the different models into the deep neural network, as they would allow the network to assign different weights to different models per output class. This improved the overall prediction accuracy, which was based on the relative prediction accuracy for each model per output class. The experiment gave us the opportunity to compare six models, which meant that the results obtained are more important than the results obtained by comparing only two models.

The experiment also concentrated on analysing the impact of balancing a data set that was initially unbalanced. We used under-sampling to remove bias from the results, and a significant improvement on the performance of all models was achieved by applying the under-sampling process. Techniques relating to data pre-processing–such as feature scaling using z-scores and converting categorical to numeric variables–were studied in detail during the experiment and subsequently applied to the data in order to improve the outcomes.

As far as the limitations of the experiment are concerned, the study was based on records relating to patients from a particular hospital, and may therefore have been biased towards the population of a specific region. Additionally, the time span used for monitoring the patients was small (5 years). To provide improved forecast results, the period of observation should be increased in order to obtain comparatively stable values, and the result of this might have an effect on predictive modelling results.

D. Summary of Analysis

The results and evaluation of the research has been discussed in this section. All six models were built on two data sets, one with biased values in terms of the target variables and one with balanced values. Cross-validation was performed using LR and RF to obtain optimal parameters in order to enhance the models’ performance. The ensemble models (RF, GB and DL) outperformed the RF, SVM, LR, GB and DL models in the prediction of mortality risk for both the balanced data set (Total Accuracy 80.65%) and the imbalanced data set (Total Accuracy 97.44%). In the ensemble model, we applied a new approach that no study has previously attempted by feeding the output probability per class from each model into the deep neural network, as this network would assign different weights to different models per output class. This improved the overall prediction accuracy in our experiment, which in turn was based on the relative prediction accuracy for each model per output class. We can therefore recommend this approach in other areas that have multiclass classification problems. The result also indicated a weaker performance of the DL model on a balanced dataset (Total Accuracy 70.89%) than on an imbalanced dataset (Total Accuracy 83.56%).

The strengths and limitations of the results concentrate on the data pre-processing techniques, which were used to improve models performance. The concluding section which follows will offer a detailed summary of the study, as well as participation and effects, and will also offer avenues for further research.

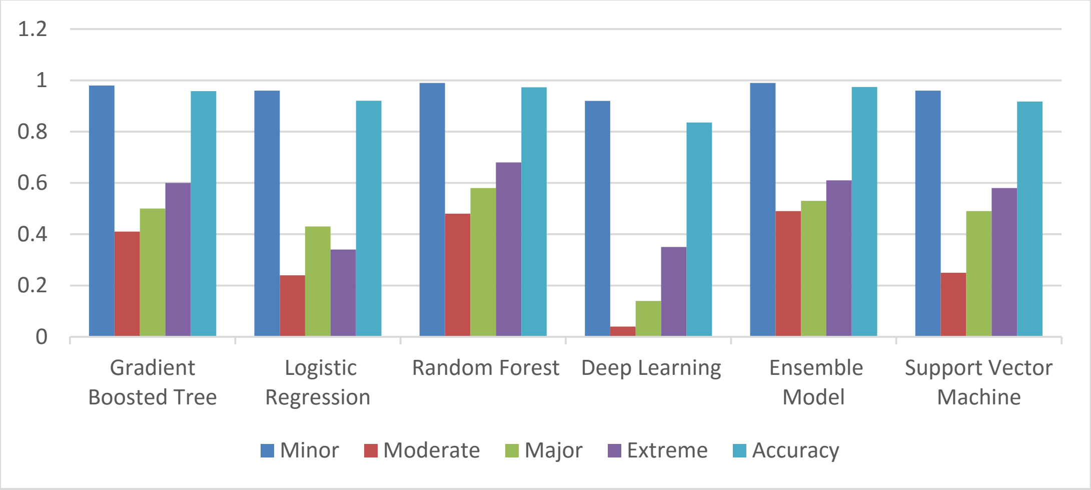


Fig. 19. Model Comparison: Imbalanced Dataset.

TABLE. III. PERFORMANCE IMBALANCED TARGET (MORTALITY RISK)

	Accuracy	**Precision**				**Recall**				**F1 Score**			
		Minor	**Moderate**	**Major**	**Extreme**	**Minor**	**Moderate**	**Major**	**Extreme**	**Minor**	**Moderate**	**Major**	**Extreme**
Gradient Boosted Tree	95.76%	1.00	0.28	0.49	0.56	0.97	0.78	0.50	0.65	0.98	0.41	0.50	0.60
Logistic Regression	92.03 %	1.00	0.15	0.37	0.22	0.93	0.58	0.52	0.71	0.96	0.24	0.43	0.34
Random Forest	97.26%	1.00	0.39	0.55	0.72	0.98	0.63	0.61	0.59	0.99	0.48	0.58	0.68
Deep Learning	83.56%	1.00	0.03	0.07	0.31	0.85	0.10	0.93	0.40	0.92	0.04	0.14	0.35
Ensemble Model	97.44%	0.99	0.40	0.59	0.55	0.99	0.63	0.48	0.70	0.99	0.49	0.53	0.61
Support Vector Machine	91.70%	1.00	0.15	0.46	0.54	0.93	0.75	0.53	0.64	0.96	0.25	0.49	0.58

TABLE. IV. PERFORMANCE BALANCED TARGET (MORTALITY RISK)

	Accuracy %	**Precision**				**Recall**				**F1 Score**			
		Minor	**Moderate**	**Major**	**Extreme**	**Minor**	**Moderate**	**Major**	**Extreme**	**Minor**	**Moderate**	**Major**	**Extreme**
Gradient Boosted Tree	76.34	1.00	0.52	0.63	0.78	0.91	0.70	0.63	0.73	0.95	0.60	0.63	0.75
Logistic Regression	71.87	0.99	0.52	0.56	0.71	0.90	0.85	0.48	0.63	0.94	0.64	0.52	0.67
Random Forest	79.09	1.00	0.67	0.65	0.77	0.92	0.67	0.67	0.82	0.96	0.67	0.66	0.80
Deep Learning	70.89	0.95	0.50	0.52	0.71	0.94	0.67	0.46	0.66	0.94	0.57	0.49	0.68
Ensemble Model	80.65	0.99	0.62	0.69	0.76	0.91	0.73	0.63	0.81	0.95	0.67	0.66	0.78
Support Vector Machine	70.99	0.93	0.44	0.57	0.86	0.86	0.67	0.67	0.58	0.89	0.53	0.62	0.69

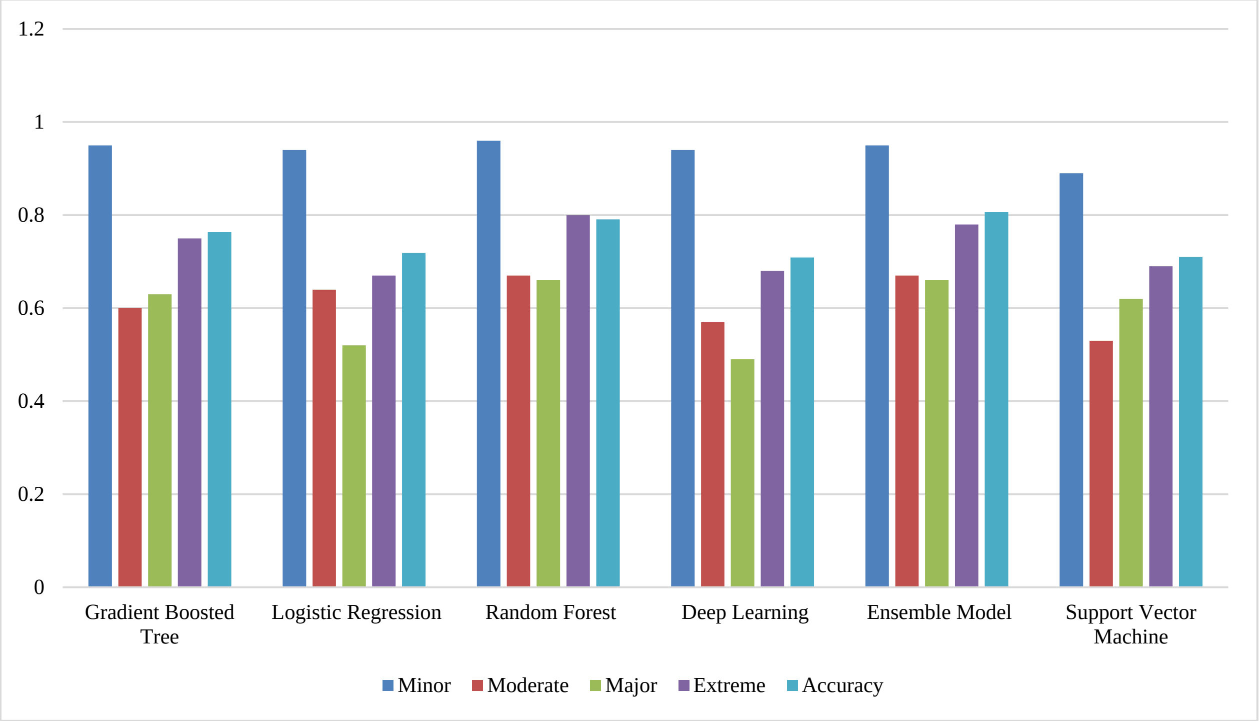


Fig. 20. Model Comparison: Balanced Dataset.

VI. Conclusion

A. Research Overview

This dissertation took the form of an investigation of multiple supervised machine learning techniques. These techniques were used to analyse factors relating to discharge details concerning infant patients. The work offered a literature review that summarised existing studies into both machine learning and mortality prediction. An experiment using six supervised classification techniques was performed in order to construct predictive mortality risk models using data that had been collected from the New York State Department of Health's state wide planning and research cooperative system over a five-year period. The first used the Support Vector Machine (SVM) technique to classify supervised learning techniques, creating a model using line or hyperplane data. Logistic Regression (LR) is a more traditional predictive technique in medical study, where the probability of multi classed occurrences is examined. The Gradient Boosting (GB) is a powerful technique for building predictive models that include weak learners, loss function and the additive model. Random Forest (RF) is a prediction algorithm which is used to run test feature data through randomly created trees. At the first connected layer, each neuron in the learning algorithm receives input from all factors from the previous layer. Finally, the Ensemble method combined several machine learning techniques (in this case DL, RF and GB) into a single predictive model in order to improve prediction levels All techniques could be used to predict mortality risks for infants using patients' information in different ways. The main purpose of the study was to measuring the model accuracy and the F1 score, and to compare the performance of each model in order to conclude which model offers the best performance in terms of prediction accuracy.

B. Problem Definition

The limitations identified in the existing literature and gaps in the research were used as motivation for the dissertation. Rinta-koski and Simo (2017) suggest that more promising methods such as SVM can be used to identify clinical features on arrival at the Neonatal Intensive Care Unit, as well as features observed during the first 72 hours of care for 598 Very Low Birth-Weight infants. However, Ahmadi et al (2017) applied Random Forest techniques to survey maternal risk factors that were associated with low birthweight neonates, using data mining on information collected from Milad Hospital to account for interactions between variables. The most commonly used algorithm to identify diseases is logical regression, so comparisons of accuracy were made between Logistic Regression, Support Vector Machine, Random Forest, Deep Learning, Gradient Boosted Tree and the Ensemble model.

The experiment was performed to empirically determine which of the six classifiers offers the better performance, giving a positive answer to the research question asked at the start of the dissertation, which was "Can we identify which features help to predict infant mortality?" No study has yet applied ensemble machine learning methods concentrating on assigning different weights to different models per output class in order to obtain a better predictive performance for infant mortality.

C. Future Work and Recommendations

This project focused only on patients from a particular hospital, and might have biased towards the population of a specific region. Further research should be conducted on monitoring and capturing more patient information from hospitals in different regions or countries, which would help build a more generalizable model. There were important variables that could not be considered in this experiment, including the mother's age, which could be useful for analyzing a new approach to create labels using three categories (18 to 28, 29 to 39 and 40 to 49) in an attempt to identify relation between the age of the mother and infant mortality risk. These data should also be collected and analysed to increase prediction accuracy.

REFERENCES

[1] J. Xu, S. L. Murphy, K. D. Kochanek, E. Arias, and D. Ph, "Mortality in the United States , 2015," no. 267, pp. 1–8, 2016.

[2] X. Kong et al., "Neonatal mortality and morbidity among infants between 24 to 31 complete weeks : a multicenter survey in China from 2013 to," BMC Pediatrics, pp. 1–8, 2016.

[3] C. E. Rodriguez-Martinez, R. Acuña-Cordero, and M. P. Sossa-Briceño, "Predictors of prolonged length of hospital stay or readmissions for acute viral lower respiratory tract infections among infants with a history of bronchopulmonary dysplasia," Journal of Medical Virology, vol. 90, no. 3, pp. 405–411, Mar. 2018.

[4] L. C. Mullany et al., "Breast-Feeding Patterns , Time to Initiation , and Mortality Risk among Newborns," no. April, pp. 599–603, 2018.

[5] A. R. Rao and D. Clarke, "An open-source framework for the interactive exploration of Big Data : applications in understanding health care," pp. 1641–1648, 2017.

[6] C. D. Smyser et al., "NeuroImage Prediction of brain maturity in infants using machine-learning algorithms," NeuroImage, vol. 136, pp. 1–9, 2016.

[7] O. Rinta-koski and S. Simo, "Gaussian process classification for prediction of in-hospital mortality among preterm infants."

[8] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME : Data Mining Methodology for Engineering Applications-A Holistic Extension to the CRISP-DM Model ScienceDirect DMME : Data Mining Methodology for Engineering Applications – A Holistic Extension to the CRISP-DM Model," 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering, vol. 96, no. July, 2018.

[9] M. Abadi et al., "TensorFlow : A System for Large-Scale Machine Learning This paper is included in the Proceedings of the TensorFlow : A system for large-scale machine learning," 2016.

[10] G. Moncecchi, "Learning scikit-learn : Machine Learning in Python."

[11] R. A. Hummer, "Black-white differences in health and mortality: A review and conceptual model," Sociological Quarterly, vol. 37, no. 1, pp. 105–125, 1996.

Framework for Digital Data Access Control from Internal Threat in the Public Sector

Haslidah Halim[1]
ICT Consultation Unit
MAMPU, Cyberjaya, Malaysia

Maryati Mohd Yusof[2]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Bangi, Malaysia

***Abstract*—Information management is one of the main challenges in the public sector because the information is often exposed to threat risks, particularly internal ones. Information theft or misuse, which is attributed to human factors, affects the reputation of public sector organizations due to the loss of public trust in the security and confidentiality of the information and personal data that are hacked by internal parties. Most studies focus on general problem solving related to internal threats instead of digital personal data protection. Therefore, this study identifies the main security control elements for personal data access in the public sector, including information security management, human resource security, operational security, access control, and compliance. A comprehensive framework is developed based on the identified security control elements and validated using a case study. Data are collected using interview, observation, and document analysis techniques. The findings contribute to the management of information system security through a systematic approach to controlling internal threats in the public sector. This framework can serve as a guideline for the public sector in managing internal threats to reduce security incidents involving unauthorized access to digital personal data.**

Keywords*—Information management; internal threats; control framework; risk; information security; personal data access

I. Introduction

Information is a critical asset in organizations. Therefore, it must be secured and protected from any modification, unauthorized use, and integrity loss [1-4]. Organization information is prioritized to ensure constant security and protection to prevent leakage to unauthorized parties. Information leakage affects organization reputation due to loss of trust in securing information from external entities.

Information systems (IS) are vulnerable to various threats, which can lead to undesired and costly consequences, including breach of data confidentiality or integrity and system unavailability [5]. Threats are attributed to internal and/or external entities. Internal threats have pervasively become a critical and serious concern in most public agencies and industries [6,7]. Most internal attacks are attributed to human factors [8-10]. According to Shu et al. [11], sensitive data leakage from computerized systems has also become a serious threat to organizational security.

Previous studies focused on solutions to general internal threats instead of internal threats pertinent to personal data [12]. Personal data are defined as processed information (partial or complete) in commercial transactions that are directly or indirectly related to individual data subjects, enabling data identification from that information or other information matching based on data related to a user. Therefore, the current work identified the main elements of security control, and these elements were utilized as the foundation of the proposed framework for internal threat control in accessing digital personal data in the public sector. The framework was developed based on a critical analysis of the literature on internal threats and of previous frameworks on policy and information security management, human resource and operational security, and access and compliance control.

This paper is divided into seven sections. Section I outlines the introduction. The literature review is discussed in Section II, and information security management and frameworks on internal threats to information security are explained in Section III. The proposed framework is described in Section IV, and the methods, results, discussion, and conclusion are presented in Sections V, VI, VII and VIII, respectively.

II. Information Security Management

Information management success is an antecedent for establishing and maintaining organization competitiveness [13]. Organization information management should be aligned with the aspect of information security to ensure information validity and accuracy. IS security in the public sector is crucial due to its role as part of critical organization infrastructure, encompassing personal or sensitive data [14]. IS in the public sectors are different from that in the private sector. IS failure in the public sector can considerably disrupt various economic and social activities and harm human life (such as failure in emergency service systems). IS security is becoming increasingly challenging due to emerging Internet-based applications, including e-commerce and various information selling services. Therefore, IS must be designed with guaranteed confidentiality, integrity, availability, authenticity, and auditability [15].

The main concern among security experts is minimizing threats from internal individuals [9, 16]. Data leakage and selling are pervasive due to hidden web site use and convenience in leaking confidential data without traces or names. Moreover, most internal attacks involve skilled and knowledgeable individuals [17].

Problems or failures in information security are normally attributed to human actions and can lead to costly losses [9, 10, 18, 30]. However, Hashem et al. [19] argued that internal

We greatly appreciate funding received from the Universiti Kebangsaan Malaysia (DIP-2016-033) and Malaysian Public Service.

threats are inevitable but can be avoided in the early stages. Therefore, a comprehensive framework for monitoring and detecting internal threats is essential for early detection.

ISO/IEC 27001/27002 is an International standard that has been widely applied in information security management. The public sector also uses this standard to protect critical information and address intrusion risks, which can lead to leakage in official government information. ISO/IEC 27001/27002 provides a framework that guides organizations in implementing the International Information Security Management Standard (ISMS) [20]. ISO/IEC 27001:2013 [21] outlines requirements for establishing, implementing, maintaining, and continuously enhancing information security management in organizational contexts. The standard also features requirements for evaluating and maintaining information security risk based on organization needs. The Malaysian government has established a cyber-security framework for the public sector [22] to provide basic guidelines that include all necessary security components that must be considered by government and public agency sectors to protect information in their cyberspace. This study employed 14 security control elements from ISO/IEC 27001:2013 [21] as bases for information security due to their inclusion in the ISMS standard. A critical analysis of previous frameworks was conducted to identify elements related to data access control.

III. Internal Threat for Information Security Framework

Three frameworks related to internal threats were identified from the literature based on their relevance, suitability, and popularity in providing solutions to problems related to such threats. These frameworks are 1) adaptive risk management and access control framework [23], 2) insider threat security architecture (ITSA) [24], and 3) management policy for access control from a system user perspective in collaborative environments [13]. These previous frameworks were compared against security control elements in ISO/IEC 27001:2013 [21] (Table I).

The framework of Baracaldo and Joshi [23] emphasizes human resource, access control, and compliance for internal users. The framework also includes risk management by considering user behavior. All security control elements in application systems are critical for minimizing internal threat risk in general and information misuse in particular.

The ITSA framework [24] is nearly complete and contains most of the security control elements in ISO/IEC 27001:2013. The framework features six security control elements, namely, information security basis, information security management, human resource security, operational security, access and compliant control for handling internal threat problems. In addition, policy elements are adopted, wherein access control is aligned with organization policy compliance. The framework acts as a control mechanism to address internal threat problems that stem from authorized system users. Moreover, ITSA framework [24] used audit elements in the monitoring process by recording audit trails, reports, and analyses in integrated business intelligence components that are appropriately aligned with compliant elements that require user activity to be recorded for security purposes.

The framework of Ahmad et al. [12] combines several security control elements, namely, information security management basis, human resource security, access and compliance control for handling internal threats. Ahmad et al. [12] provide autonomy to data owners by enabling their direct involvement in managing policy for data access. All security control elements in application systems are critical for minimizing risk caused by unauthorized use or information theft by authorized users.

All three frameworks use the same security control elements, namely, information security basis, information security management, human resource security source, access and compliance control for handling internal threat problems in organizations. Only the ITSA framework [24] adopts operational security elements whereas that of Baracaldo and Joshi [23] includes risk management elements.

TABLE. I. Framework Comparison based on Security Control Elements in ISO/IEC 27001:2013

	Framework		
Security elements ISO/IEC 27001:2013	**Baracaldo and Joshi [23]**	**ITSA [24]**	**Ahmad et al. [13]**
Information security basis	Available	Available	Available
Information security management	Risk management and control access	Audit Security (threat prevention)	Application and database management
Asset management			
Human resource security	Control via monitoring, context, and trust modules	Authorization of access control	Access control on user and data owner
Physical and environmental security			
Cryptography			
Operational security		Audit trail logs Reporting and analysis	
Communication security			
Access control	Available	Available	Role-based access control
IS acquisition, development, and maintenance			
Management of security incidents			
Supplier relationship			
Service continuity management			
Compliance	Compliance Policy enforcement	Policy compliance Management order Rules and regulations	Policy implementation

The literature on internal threats was reviewed to identify additional elements for security control in handling internal threat problems. Security documents regarding government ICT and Personal Data Protection Act 2010 [25] were also analyzed considering the study scope in developing a framework for internal threats to protect digital personal data for public service use. Six security control elements and three additional elements were identified from the literature (Table II). In short, the six main security control elements, namely, information security basis, information security management (ISM), human resource security, operational security, access and compliance control, are critical in handling internal threats for accessing the digital personal data of the public sector. Security control implementation is supported by three additional elements in PDPA, namely, risk management, cryptography, and access principle. All six security control elements and the three additional elements served as the foundation for developing the proposed framework for internal threat control from personal digital data access (PDDAITC) in the public sector.

The framework by Ahmad et al. [12] was adopted and extended by adding five new elements, namely, internal threat control, security policy control, application control, database control, and security control. The comparison of the proposed framework and that of Ahmad et al. [12] is shown in Table III.

TABLE. II. SECURITY CONTROL AND ADDITIONAL ELEMENTS

	Elements	**References**
Security control elements		
1	Information security policy	[15,23,24,26,33]
2	Information security management	[14,15,23,24,26,31]
3	Human resource security control	[15-17,18,19,23,24,26-28]
4	Operational security	[23,24,26]
5	Access control	[6,23,24,26,33]
6	Compliance	[23,24,26,30,32,33]
Additional elements of security control		
1	Risk management	[3,15,22,25,31,33]
2	Cryptography	[2,11]
3	PDPA – access principle	[25,33]

TABLE. III. COMPARISON OF THE PROPOSED FRAMEWORK WITH THAT OF AHMAD ET AL. [12]

Ahmad et al. [13] Framework	**Proposed Framework (PDDAITC)**	
User Data admin/ owner	Internal threat source	User Data administrator/ owner
Server policy	*Application control	Access control (id, password, cryptography) System authorization control
Database	*****Database control -data owner autonomy	Access control (id, password, cryptography) Audit trail Encryption
Policy - Data owner determines data access policy	*Security policy control	ISM ISMS standard Risk management PDPA
	*Internal threat control	Access control policy Guideline/SOP ICT security policy Operational security
	*Security control	Compliance

IV. PROPOSED FRAMEWORK

The proposed framework was developed using a holistic approach by adapting the framework of Ahmad et al. [13] and conducting a critical analysis of the literature related to internal threats and control elements (Fig. 1). Six security control elements and three additional elements are classified as security components, namely, internal threat control, security policy control, application control, database control, and security control. All proposed security elements are featured in ISO/IEC 27001:2013 [21] security elements, which are also encompassed in the reviewed frameworks [12, 23, 24]. Users and data owners, which are classified as sources of internal threats, must comply with all security controls.

Before access is granted, data owners must comply with security policy control, which comprises ISM, risk management, standard organization ISMS, and PDPA. Data owners must also adhere to database control security, which includes access control element, audit trail, and encryption. Access control authorization for data owners is only verified through user ID, password, and cryptography. Data owners are granted autonomy on their data based on the access principle in PDPA. Users and data owners must comply with all security control categories in the proposed framework to secure organization control from any undesired security incident. The following security control elements and their implementation approaches are outlined.

1) All appointed officers, staff, and suppliers must comply with the information security policy. Violation of information security can affect confidentiality, integrity, and information availability.

2) Security control of human resource is a critical element in controlling internal threats in organizations. The security filter in a non-disclosure agreement (NDA) is a document that must be signed by suppliers and recorded in a system. All officers, staff, and appointed suppliers must also sign and comply with the security policy of organization ICT.

3) Information security management is an important element that must be considered by organizations because any leakage or unauthorized use of information can affect organizational reputation. Undesired incidents are controlled and prevented through ISMS organization security policy.

4) Operational security - Application and database operation are monitored based on SOP or guidelines to ensure the smooth daily operation and avoid undesired security incidents.

5) Application and database access control is determined in accordance with the role, job scope and work duration of officers, staff, and suppliers. Access cancellation for relocated, retired, or terminated staff must be immediately activated to minimize risks of security threat toward the organization.

6) Compliance - Security or guideline policy and SOP must be fulfilled to ensure enforcement on appointed officers, staff, or suppliers. Follow-up action must be executed against security incidents pertinent to security information.

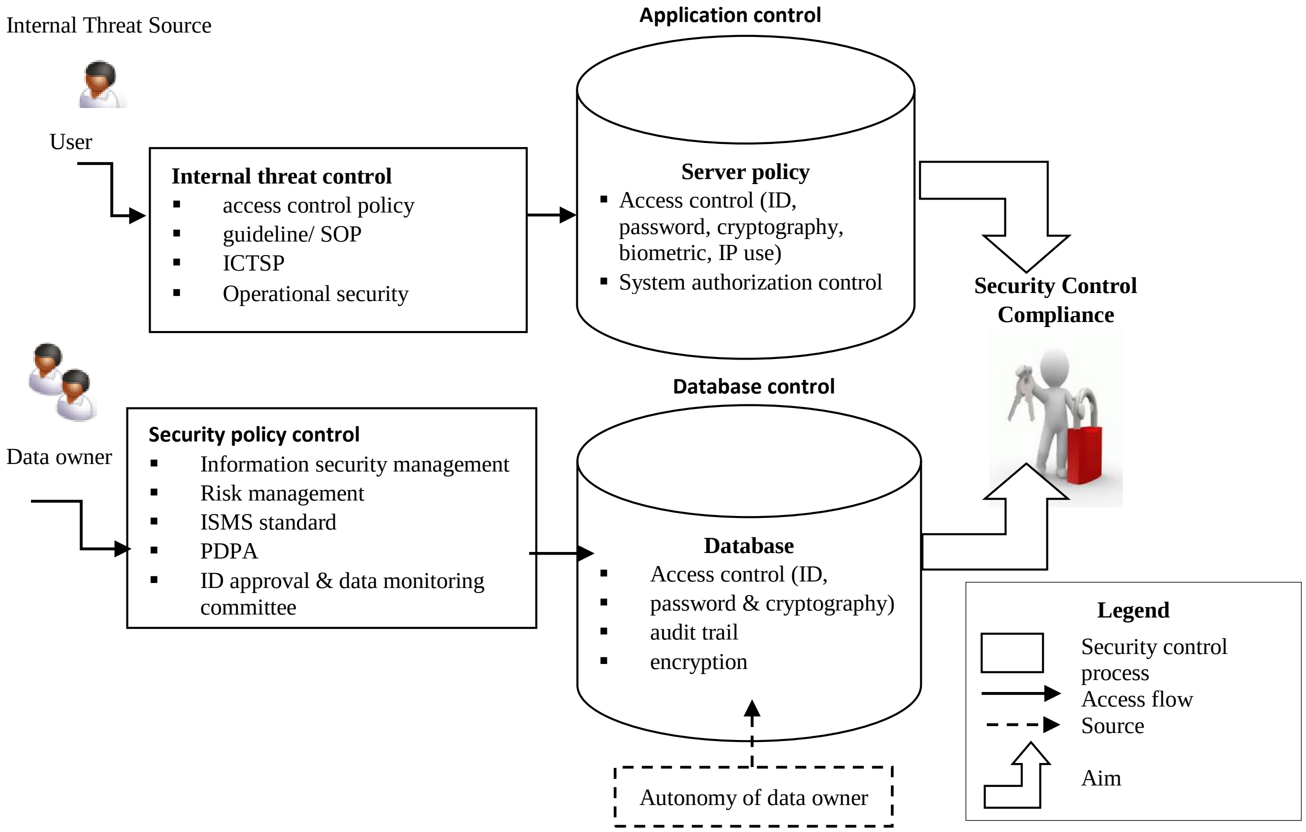


Fig. 1. Framework for Internal Threat Control for Digital Access Control in Public Sector.

The following are additional elements that support the implementation of security control:

1) Risk management is identified and monitored through the enforcement of information security policy on appointed officers, staff, or suppliers to reduce intentional or unintentional threat risk.

2) Cryptography is implemented on applications and databases to increase the security of information stored in the organization.

3) Access principle (seventh PDPA principle) provides autonomy for data owners to update data and execute legal action in case of personal data misuse.

V. Method

This study employed a qualitative case study to understand and comprehensively describe the social phenomenon under study [29]. Data were collected using interview, document analysis, and observation techniques. Security control elements and the proposed framework were validated based on a case of a public sector agency in Malaysia, known as the National Registration Department (NRD). One-on-one, face-to-face interviews were conducted with seven expert informants at the NRD. These individuals were directly involved with ICT applications (mainly those involving personal data and their sharing with other agencies) and database security of the public sector. The informants and their profiles are listed in Table IV.

The interview agenda was developed based on the research questions and the proposed framework. The agenda was pilot tested with an informant to evaluate the appropriateness of the interview questions. During the interview sessions, the experts were briefed on security control elements. The validated security control elements were refined in the initial framework. The implementation and enforcement processes of the access control system in the NRD office were directly observed. These processes included ID card use among officers and staff, password and biometric use on application systems (particularly at the NRD main counter), and password use for logging on to officer and staff computers. Access control technology, either physical or computerized systems, was also monitored. Important records were also analyzed, they include government documents pertinent to ICT security of public sector, PDPA 2010 (Act 709), ICT security policy for NRD and ISO/IEC 27001:2013, Identification Card System (one of the main systems at NRD counter), and Agency Link-up System (ALIS), one of the systems used for sharing data with statutory body. Documentation for both systems was analyzed from access control implemented on the user who accessed the systems. Data were audio- and hand-recorded and transcribed. They were then organized based on themes and tabulated for subsequent analysis, discussion, and conclusion.

TABLE. IV. INFORMANT LIST

Expert code	Position	Work duration (year)	Expertise (year)
1	Assistant Senior Director (ASD)	12	Application project management (4) ICT security (8)
2	ASD	12	Application project management and development (12)
3	ASD	10	Database (10)
4	ASD	14	Application project management and development (14)
5	ASD	11	Application project management and application (11) ICT security (Internal auditor of ISMS in NRD (5))
6	ASD	14	Application project management (6) Database (5) ICT security (2)
7	ASD	14	Application project management and application (14)

VI. RESULTS

NRD was selected as the case study based on its profile as a government agency that stores data of all Malaysian residents and its role as the main reference for all government, private, and statutory body agencies. NRD also systematically manages information security by obtaining ISO/IEC 27001:2005 and ISO/IEC 27001:2013 certifications. Overall, all the interviewed experts understood the meaning and importance of internal threats at NRD. According to Experts 1, 2, 3, and 4, internal threats include information or technology misuse and document forgery from their own organizations, units, departments, or ministries involving appointed officers, staff, suppliers, or other entities directly involved with NRD.

In addition to ISMS certifications, NRD has also established various procedures and SOPs to ensure ICT security. Numerous experts have confirmed the occurrence of internal threat incidents involving personal data at NRD. According to Expert 4, data from ALIS system were misused and disseminated to an unauthorized third party. Consequently, Expert 4 was interrogated by the Integrity Commission Agency. Furthermore, a server log checks indicated attempts to obtain additional personal data (Expert 3). NRD has strengthened its security mechanism by ensuring non-recurring incidents. The agency also provides policy, guidelines, and briefings on security awareness to all appointed officers, staff, and suppliers. NRD adopts several methods or mechanisms to address internal threats to application and database access involving personal data, namely, internal threats, applications, and database control.

NRD has established access control policy and SOP as procedures to prevent unauthorized use of data or information. A committee was established to approve user ID application for ALIS system. In the case of ID misuse, prevention mechanisms, such as the denial of access and cancellation of IDs and passwords, would immediately take effect. “NRD has a detail access control policy. New or old staff can obtain or remove access upon his relocation or retirement through a specific method” (Expert 7). All experts mentioned background investigations on staff and suppliers involved in ICT projects at NRD; these investigations are conducted to control physical access (human security) to ICT assets. According to Expert 1, “suppliers are obligated to provide service admission letter and Official Confidentiality Act that must be renewed yearly and fill out security clearance under Chief Government Security Officer and obtain approval from NRD security policy”. Experts 2 and 5 supported this statement by stating that NRD practices the compliance principle of ISMS, which requires appointed staff and suppliers to fill out NDA forms.

NRD is yet to establish a specific security policy pertinent to internal threats. However, the agency referred to the highest level and a general ICT security policy [26] in addition to the following: 1) the other policies of the department, including access control procedures on applications, databases, procedure IDs, passwords, and data sharing between the agency and the private sector; 2) meetings and a committee for monitoring access control on issues and NRD problems. These endeavors show the remarkable commitment of NRD management to information security. Furthermore, most of the experts understood the basic PDPA 2010 that was applied to the ALIS system, which involves data sharing with statutory bodies. All appointed officers, staff, and suppliers must comply with all security policies at NRD and sign the recommendation of ICTSP [27] of NRD and the official government act.

Continuous monitoring is implemented to prevent security incidents at NRD, particularly those involving information misuse or leakage. According to Expert 1, risk management has also become the main agenda at NRD because of its inclusion in ISMS and yearly implementation. The other experts also acknowledged the importance of implementing and monitoring risk that is pertinent to internal threats. Expert 1 explained, “Any risk encounter will be mitigated based on the identification of appropriate method and solution alternatives.”

The experts perceived the importance of application control at NRD, which is implemented by ensuring authorization of user access to applications and information. Security and application control are enhanced through system verification, including IDs and passwords, biometrics, and procedures and SOPs pertinent to operational security applications, to ensure smooth daily operation. Despite the importance of cryptography, its application remains costly for the limited budget of NRD.

NRD is the main agency that manages the sensitive data of Malaysian citizens. Therefore, NRD prioritizes database control and has established a specific procedure for its control. The agency tracks database audit trails in cases of problems related to database access. According to Expert 3, “…monthly checking is performed based on the procedure for registering and canceling database access.” An audit trail is a critical component of ISMS. The experts agreed on database cryptography at NRD, as implied by Expert 1, “ISMS NRD is audited by internal and external auditors, involving detailed audit trail check. Auditor check access for active ID and immediate access cancellation for inactive ID for relocated or retired NRD staff.”

All the experts generally agreed with security control elements in the proposed framework based on their clarity and appropriateness to the study context. They also suggested additional elements, namely, ID approval committees and data monitoring data in security policy control and biometrics for IP use, to control user access in application control components.

VII. Discussion

Practice-based research in information security has been advocated to provide insights in actual conduct, challenges, and mitigation approaches implemented in organizations [30]. The validation of the proposed framework indicated the importance on six core and three complementary security controls (information security policy; human resource security control; information security management; operational security; access control; and compliance; risk management; cryptography; and PDPA access principles). The security controls indicated the importance of technical and socio-technical factors, particularly human factors, in ensuring a comprehensive and effective mechanism to protect information. Other studies also corroborate with the significant role of socio-technical aspects in information security [15, 30].

NRD has cultivated a strong security culture through various controls, communication, enforcement mechanisms in a timely, prospective and retrospective, continuous, and comprehensive manner. Despite the absence of a security policy for internal threat, NRD proactively referred to other general, applicable security policies. This is in line with a related study in the Swedish public sector that has no security policy but adopted other practiced information security management approaches [31]. The study associated this workaround as the attitude of knowing how to be “good enough” that adopt, adapt and enhance any available and applicable security measures to the relevant context.

The findings also show that the control mechanism is only effective with continuous monitoring, implementation, compliance, and cooperation from all stakeholders. The concepts are closely related to those of governance, risk, and compliance (GRC) [31, 32].

VIII. Conclusion

This study contributes to the internal threat management discipline, particularly for personal digital data in the public sector. The proposed framework can guide users, specifically managers and officers involved in application, database, and ICT security in the public sector, in protecting organization information from threats or security incidents caused by internal threats. Therefore, risk incidents can be prevented in their early stages to enhance information security. The findings also contribute to organization practice in information security pertinent to access control, a critical security domain in collaborative work environments and various computerized or physical platforms. The framework can serve as a guideline to ensure systematic management of threat control for secured personal data access.

The paper has a number of limitations. Although the scope is limited to one agency, the framework can also be applied in any ministry or public agency because it involves general features for internal threats involving personal digital data. The framework can positively influence government efforts at the information security level to protect personal digital data from any security incident, which can affect data integrity and public sector reputation.

Further research can be conducted on a wider context of local or international public agencies to obtain richer, holistic, and context-specific overview and lesson learned on the subject matter. The research scope can also be extended to information security services for network security monitoring and government security incident management purposes. Further work on all related scope, namely application, database, network security management, and security incident may enhance ICT security in the public sector.

The research scope is also limited to internal threat for protecting digital personal data access only. Therefore, further research on offline data can be conducted as these data are also vulnerable to an internal threat risk. Similarly, the work scope can also be extended to external threat perspective. More work can also be performed on control elements beyond the proposed framework that apply to internal threat in the public sector. The framework components need to be enhanced accordingly in the future based on the requirement, technology, and organizational changes.

The comprehensiveness and appropriateness of the proposed framework in addressing internal threats of personal digital data access in the public sector were validated. The comprehensive framework is applicable in supporting public sector environment and practice in managing internal threat systematically. Based on security policy and procedure and ISMS practice in NRD, the six core security elements are capable of mitigating internal threat for digital personal data access.

Acknowledgment

We thank all study participants from NRD for their collaboration.

References

[1] G. Pavlov and J. Karakaneva. “Information security management system in organization”. Trakia J Sci, vol. 9, no. 4, pp.20-25, 2011.

[2] M.A. Mizhera, R. Sulaiman and A. M.A. Abdalla. “An improved simple flexible cryptosystem for 3D objects with texture maps and 2D images”. J Inf Sec Appl vol 47, pp. 390-409, August 2019.

[3] A. Alwi and K A. Zainol Ariffin. “Information Security Risk Assessment for the Malaysian Aeronautical Information Management System” Cyber Resilience Conference. Putrajaya, Malaysia, November 2018.

[4] A. H. Kashmar, A.K. Hassn and E.S. Ismail. “Hybrid chaotic keystream generation (HCKG) for symmetric image encryption”, J Theor Appl Inf Tech, vol. 97, no. 3, pp. 984-993 1 Feb 2019.

[5] M. Jouinia, L.B.A Rabaia and A. Ben Aissab. “Classification of security threats in information systems”. 5th Intl Conf Ambient Systems, Networks and Technologies, Hasselt, Belgium. June 2014.

[6] P.A. Legg, O. Buckley, M. Goldsmith and S. Creese, S. “Caught in the act of an insider attack: detection and assessment of insider threat”. IEEE Int Symposium on Technologies for Homeland Security, Waltham, USA, 2015.

[7] J. Eggenschwiler, I. Agrafiotis and J.R.C. Nurse. “Insider threat response and recovery strategies in financial services firms”. Comput Fraud Security, vol. 11, pp.12-19, 2016.

[8] A. Price and Y.B. Choi, “Human factors in information security”. Int J Comput Inf Tech, vol. 4, no. 5, pp. 833-847. Sep, 2015.

[9] Wan Ismail, W.B. and Yusof, M.M. “Mitigation strategies for unintentional insider threats on information leaks”, Int J Secur Appl, vol. 12, no.1, pp. 37-46I, 2018.

[10] Wan Ismail, W.H.B. and Yusof, M.M. “Assessing data leakage prevention for data-in-use”, Pacific Asia Conference on Information Systems, Langkawi, Malaysia, July 2017.

[11] X. Shu, J. Zhang, D. Yao, S. Membe, and W.C. Feng. “Fast detection of transformed data leaks.” IEEE Trans Inf Forensics Secur, vol. 11, no. 3, pp. 528-542. 2016.

[12] S. Ahmad, S.Z.Z Abidin, N. Omar and S. Reiff-Marganiec. “Managing access control policy from end user perspective in collaborative environment”. IEEE Conference on Open Systems, Subang, Malaysia. October 2014.

[13] W.I.W. Sulaiman and M.H. Mambob, “Significance of communication satisfaction model in the context of information management of public sector” Malay J Comm, vol. 30, no. 1, pp. 97-115, 2014.

[14] E. Loukis, and D. Spinellis, “IS security in the Greek public sector”. Inf Manage Comput Secur, vol. 9, no.1, pp. 21-31, 2001.

[15] A.I. Al-Darwish, and P. Choe. “Application of a human factors-integrated information security framework to an oil and gas organization”. Adv Intell Syst Comput. Vol. 1018, pp. 731-736, 2020.

[16] N. Elmrabit, S.H.Yang, and L. Yang, “Insider threats in information security categories and approaches”. 21st Int Conf on Automation and Computing: Automation, Computing and Manufacturing for New Economic Growth, Glasgow, UK. September 2015.

[17] A Sanzgiri, “Classification of insider threat detection techniques”. Proceedings of the 11th Annual Cyber and Information Security Research Conference, Tennessee, USA, April,2016.

[18] S. Soltanmohammadi, S.Asadi, and N.Ithnin. “Main human factors affecting information system security”. Interdiscip J Contem Res Bus vol. 5, no. 7, pp. 329-354, 2013.

[19] Y. Hashem, H. Takabi, M. Ghasemigol, and R. Andtu, “Inside the mind of the insider: towards insider threat detection using psychophysiological signals”. J Internet Serv Inf Secur, vol. 6, no. 1, pp. 20-36, 2016.

[20] Z. Mukhtar and K. Ahmad. “Internal threat control framework based on information security management system. J Theor Appl Inf Tech, vol. 70, no. 2, pp. 316–323, 2014.

[21] ISO/IEC 27001:2013. 2013. International Standard. ISO 27001 “Information technology - security techniques - information security management systems–requirements”. https://www.bsigroup.com/LocalFiles/en-GB/iso-iec-27001/resources/BSI-ISO27001-transition-guide-UK-EN-pdf.pdf.

[22] Malaysian Administrative Modernisation and Management Planning Unit (MAMPU), “Public sector computer security framework” (PSCSF version 1.0). http://www.mampu.gov.my/images/pengumuman/PSCSF-Versi-1-April-2016-BM.pdf, April, 2016.

[23] N. Baracaldo and J. Joshi. “An adaptive risk management and access control framework to mitigate insider threats”, Comput Secur, vol. 39, pp.237-254, 2013.

[24] G. Jabbour, and D.A. Menasce, “The insider threat security architecture: A framework for an integrated, inseparable, and uninterrupted self-protection mechanism”. Int Conference on Computational Sc and Eng Vancouver, Canada, pp. 244-251. August, 2009.

[25] Personal data protection department, Malaysia Ministry of Communication and Multimedia. Personal Data Protection Act 2010 (Act 709). Available http://www.pdp.gov.my/index.php/my/akta-709/akta-perlindungan-data-peribadi-2010 2010.

[26] ICT security policy(ICTSP) National Registration Department. MS-NRD-SM-PP-01, 5 February 2015.

[27] A. Munir, L. Sharif, M. Kabir and M. Al-Maimani. “Human errors in information security”. Int J Adv Trends Comput Sc Eng, vol. 1, no.3, August, 2012.

[28] F.L. Greitzer and R.E. Hohimer, “Modeling human behavior to anticipate insider attacks.” J Strategic Secur, vol. 4, no. 2, pp. 25-48. 2011.

[29] R. K. Yin, “Case study research: design and methods. essential guide to qualitative methods in organizational research, Thousand Oaks: Sage Publications, 2014.

[30] H.A. Hamid, M.M. Yusof, N.R.S.M. Dali, “The influence of security control management and social factors in deterring security misbehavior” Int J Recent Technol Eng, vol 8, no 1, pp 144-150, June 2019.

[31] E. Bergström, M. Lundgren and Å. Ericson, “Revisiting information security risk management challenges: a practice perspective” Inf Comput Secur vol. 27 no. 3, pp. 358-372, 2019.

[32] C. Sillaber, A.Mussmann,and, R. Breu, “Experience: Data and information quality challenges in governance, risk, and compliance management” J Data Inf Qual, vol. 11, no. 2, 2019.

[33] Personal data protection commission Singapore and Privacy Commissioner for Personal Data, Hong Kong. “Guide to data protection by design, for ICT systems”, 2019. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Data-Protection-by-Design-for-ICT-Systems-(310519).pdf.

Improving Usable-Security of Web based Healthcare Management System through Fuzzy AHP

Nawaf Rasheed Alharbe
Department of Computer Science and Information, Community College, Badr, Taibah University, KSA

***Abstract*—Security is an important concern in web application development that is getting massive consideration from academic and IT industry. In addition, due to big share in web based healthcare management system, usable-security is in much demand. However, identifying and choosing the accurate model for improving usable-security of web based healthcare management system is becoming more challenging for practitioners. Usable-security factors contribute a noteworthy role while integrating application security and application usability during development of healthcare management system. Every factor has its own significance while incorporating usable-security during development of healthcare management system. This is based upon user's demand and sensitivity of patient information. Hence, practitioners want to know about the significance of each factor when they are developing a web application to satisfy the user's request. Author of this article measured the usable-security of web based healthcare management system by using Fuzzy Analytic Hierarchy Process (Fuzzy AHP). Also, the impact of each factors of usable-security for web based healthcare management system has been given. This will help the practitioners to improve application usability of security while designing web based healthcare management system.**

Keywords*—Application security; application usability; usable-security; fuzzy analytic hierarchy process

I. Introduction

The Government of Saudi Arabia has given a high priority to providing the best practice in patients' care. A lot of studies are presented, trying to recognize and categorize the techniques in which the security of web application can be improved [1, 2]. Further, there is always been a hole between theory and practice which is hard to fill entirely, the problem can be reduced by founding a mutual methodology to increase the accessibility of outcomes. In this contribution, author has made an effort to improve accessible hypothetical research for quantifying usable-security. Usability is an energetic factor of security for web based healthcare management system. To achieve usable-security in web application during development, identification of security as well as usability factors are suitable [3, 4]. Therefore, practitioners need to understand how to relate security factors with those of usability and evaluate the impact of these factors for increasing security of web based healthcare management system.

Assessment of usable-security factors is essential to confirm it [5, 6]. Web application development has considerable potential to support "greening through IT"—that is, making civilizations more environmentally sustainable via IT interventions. To draw attention to such issues in web application development engineering, we argue that usable-security must be treated as a first class quality alongside other important and precarious attributes such as safety, security, reliability, usability, and efficiency. Results of assessment procedure may allow decision makers to make suitable decision and initiate appropriate action [7, 8]. However, to be able to take correct action, decision makers should not only know the security and usability factors but also their mapping. Hence, Fuzzy Analytic Hierarchy Process (Fuzzy AHP) is used in this contribution for prioritizing factors. To address usable-security issues of web based healthcare management system, prioritization of the factors is a critical procedure.

Rest of the paper is organized as follows: in Section 2, needs and importance is discussed. Priority assessment of usable-security factors is calculated in Section 3. Finally, significance and conclusion are given in Sections 4 and 5.

II. Needs and Importance

Plenty of research has been done in the field of selecting and ordering factors of security with fuzzy analytic hierarchy process [9, 10]. Alenezi et al. in 2019 prioritized usable-security attributes using Fuzzy AHP technique But tiny research has been completed for prioritizing security factors that affect usability of security and balancing their trade-offs with respect to healthcare management. Success of security technology largely depends on user acceptance [11, 12]. It is essential to measure usable-security factors during development of web based healthcare management system. Results assessment of usable-security factors should be analyzed deeply so that it can be used to enhance usability of secure web application [13, 14]. The analysis of prioritization is done using Fuzzy AHP, which is a type of Multi-Criteria Decision Analysis (MCDA) [15, 16]. MCDA contributes a vital role for acting numerous inconsistent estimation objects [17, 18]. MCDA methods are mainly distributed in the three classifications including objectives, alternative weights and their ranks.

Analytic Hierarchy Process (AHP) is measured for evaluating a judgment in set, but numerous practitioners have suggested that Fuzzy AHP is additional valued to deliver crisp decisions with their weightages too [19, 20]. In addition, it has been a significant tool that is widely used to complete priority examination and approved by decision makers. For paradigm a hierarchy of factors giving to their significance, AHP is functioning with decision input from a group of decision makers [21, 22]. To deal with the doubts and ambiguity of practitioner's judgment, the author took Fuzzy AHP [23-25].

Further, it is a hybrid technique of fuzzy set theory and AHP. In this contribution a manner for estimation of usable-security through Fuzzy AHP has been presented. For gathering data author has taken 101 practitioner's decisions. With the help of the inputs of practitioner's decisions, this paper estimates the importance of usable-security factors in terms of their weight and ranks. Based on the results, usable-security improvement policies are identified and selected to moderate and manage usable-security of web based healthcare management system in future.

III. Measuring Usable-Security Attributes

Usable-security factors are commonly a qualitative quantity. It is a process to evaluate usable-security factors quantitatively. Further, weightages and ranks of usable-security factors contribute an important role for extremely secure design of web based healthcare management system. Usable-security factors prioritization for the necessity of usable web based healthcare management system is a MCDM problem [8, 9]. This set of criteria regularly varies in the amount of prominence. There have been numerous methods or tools for answering this kind of problem including AHP method and numerous other methods, in which AHP has been a method that is broadly used and approved by practitioners to aid in priority analysis [10, 11].

This section discusses the methodology for deriving weightages of usable-security factors to manage these usable-security factors during security design process. Priority of usable-security factors should be decided before the designing phase. And also, during the execution, security practitioners should have knowledge of the important usable-security factors identified and classified before it can make any severe security issue [4]. Ranking and weightages of these factors are evaluated using Fuzzy AHP technique. Further, Fuzzy AHP is capable for controlling ambiguous judgment given by the practitioners. It is also helpful in converting linguistic inputs into numerical outputs, which is further helpful to prioritize these factors [8, 9]. The weightes and ranks of usable-security factors may be helpful to developers for selection of the development guidelines. In addition, these guidelines are essential to maintain the confidentiality, integrity, and availability (CIA) for usable-security. Fig. 1 discusses the different security factors of web based healthcare management system that are related to usability.

The hierarchical structure of usable-security factors is presented in Fig. 1. The factors have been identified through a comprehensive literature review and practitioners' opinions. The usable-security factors that have been considered in this contribution have already been discussed with their impact on usability [8]. For integrating usability to security, essential security usability factors that may enhance security of web based healthcare management system design have been considered in this section. The present contribution aims to determine priority of security factors affecting usability of web based healthcare management system. For this aim a questionnaire is prepared from [5]. Thus, it is required to have a group of experienced practitioners working in area of security to answer the questionnaires. For evaluating the weightages of usable-security factors form practitioner's opinion, Triangular fuzzy numbers (TFNs) equations have been used which is shown in equations (1)-(3). TFNs [η_{ij}] are established as the following:

$$\eta_{ij} = (l_{ij}, m_{ij}, h_{ij}) \qquad \ldots\ldots.(1)$$

$$where\ l_{ij} \leq\ m_{ij} \leq h$$

$$l_{ij} =\ min(J_{ijd})) \qquad (2)$$

$$m_{ij} = (J_{ij1}, J_{ij2}, J_{ij3})^{\frac{1}{x}} \qquad (3)$$

$$and\ h_{ij} =\ max(J_{ijd})$$

Where, J_{ijk} indicates the relative significance of the values F_i and F_j specified by practitioner k and i and j indicates a pair of conditions being refereed by practitioners. F_{ij} represents TFN for the comparison between criteria F_i and $F_{j..}$i.e. F_i- F_j. Comparison between criteria F_j and F_i is the reverse of F_i and F_j. Value m_{ij} is estimated based on the geometric mean of practitioner's scores. After getting the TFNs value, a fuzzy pair-wise comparison matrix is recognized in the form of n x n matrix and is shown in Table I.

The size of the comparison matrix is 6x6, the size of the group to fulfill an acceptability of consistency is 101 practitioners [8]. Practitioners of this assessment include academicians and software developers having knowledge in web application security. Sample of questionnaire is taken from [8]. After qualitative assessment, pair-wise comparisons are prepared quantitatively. The matrix prepared by the researchers after evaluating judgments of practitioners is shown in Table II.

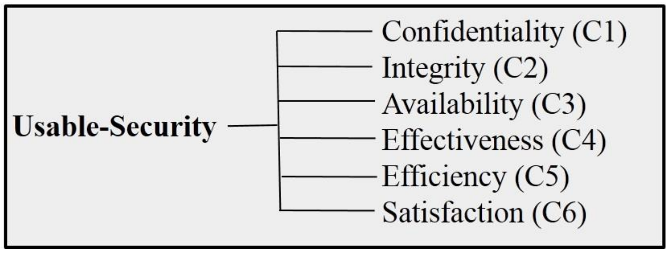


Fig. 1. Hierarchy Model for usable-Security.

TABLE. I. Example of Fuzzy Pair-Wise Comparison Matrix

		Attribute 1	Attribute 2	Attribute 3	Attribute 4	………..	Attribute n
	Attribute 1	(1,1,1)	F_{12}	F_{13}	F_{14}	………..	F_{1n}
	Attribute 2	F_{21}	(1,1,1)	F_{23}	F_{24}	………..	F_{2n}
η_{i} j =	Attribute 3	F_{31}		(1,1,1)		………..	
	Attribute 4	F_{41}			(1,1,1)	………..	
	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .	. . . .
	Attribute n	F_{n1}	F_{n2}	F_{n3}	F_{n4}	………..	(1,1,1)

TABLE. II. FUZZY PAIR-WISE COMPARISON MATRIX

	Confidentiality (C1)	**Integrity (C2)**	**Availability (C3)**	**Effectiveness (C4)**	**Efficiency (C5)**	**Satisfaction (C6)**
Confidentiality (C1)	1,1,1	1.0, 1.5, 1.9	0.5, 0.6, 1.0	0.4, 0.6, 1.0	0.2, 0.3, 0.4	0.3, 0.5, 0.9
Integrity (C2)	-	1,1,1	0.6, 0.7, 0.8	0.3, 0.4, 0.6	0.3, 0.4, 0.5	0.2, 0.2, 0.3
Availability (C3)	-	-	1,1,1	1.0, 1.3, 1.6	0.3, 0.4, 0.8	0.8, 0.9, 1.0
Effectiveness (C4)	-	-	-	1,1,1	0.5, 0.9, 1.6	0.6, 1.1, 1.7
Efficiency (C5)	-	-	-	-	1,1,1	0.4, 0.6, 1.2
Satisfaction (C6)	-	-	-	-	-	1,1,1

After it, defuzzification process is achieved for producing a measurable assessment based on the calculation of TFNs values. This has been derived from [9, 10] as formulated in equation (4), also known as the alpha cut method. Alpha threshold value is any value taken from scale of 0 to 1. For this research work, alpha threshold value has been taken as 0.5. The set $\mu_{\alpha,\beta}$ is called a strong alpha-cut set if it contains of all the fundamentals of a fuzzy set whose membership functions have principles strictly better than a quantified value. Equation (4) shows the general form of alpha cut.

$$\mu_{\alpha,\beta}(\eta_{ij})= [\beta.\eta_{\alpha}(l_{ij})+ (1-\beta).\ \eta_{\alpha}(h_{ij})] \quad (4)$$

where $0 \leq \alpha \leq 1$ *and* $0 \leq \beta \leq 1$

such that,

$$\eta_{\alpha}(l_{ij})= (m_{ij}- l_{ij}).\alpha+l_{ij} \quad (5)$$

$$\eta_{\alpha}(h_{ij})=h_{ij}- (h_{ij}- m_{ij}).\alpha \quad (6)$$

α and β in given equations are used for views of practitioners. By using equation (4) with α and β at 0.5, the result is shown in Table III. Because, the values of α and β varies between 0 and 1, the value of α and β is based on 50-50 chances.

TABLE. III. DEFUZZIFIED PAIR-WISE COMPARISON MATRIX

	Confidentiality (C1)	**Integrity (C2)**	**Availability (C3)**	**Effectiveness (C4)**	**Efficiency (C5)**	**Satisfaction (C6)**
Confidentiality (C1)	1	1.49	0.69	0.64	0.30	0.53
Integrity (C2)	0.67	1	0.68	0.41	0.37	0.20
Availability (C3)	1.45	1.48	1	1.30	0.49	0.85
Effectiveness (C4)	1.56	2.42	0.77	1	0.97	1.10
Efficiency (C5)	3.30	2.69	2.03	1.04	1	0.72
Satisfaction (C6)	1.90	4.92	1.17	0.91	1.39	1

The following stage is to calculate the eigenvalue and eigenvector. The purpose of computing the eigenvector is to estimate the weights of specific factor. Author assumed that μ signifies the eigenvector while λ signifies the eigenvalue of fuzzy pair-wise comparison matrix $\eta_{ij.}$ Equation (7) is based on the linear transformation of vectors, where *I* represent the unitary matrix.

$$[\mu_{\alpha,\beta}(\eta_{ij})- \lambda I].\ \mu=0 \quad (7)$$

The combined weightges and percentage is given in Table IV. In actual scenario, there are various usable-security attributes, which are present in web based healthcare management system development process [8-10, 12]. In this research, only six usable-security factors have been identified as well as prioritized. The hierarchy for these factors affecting usability is established and their weightage is calculated through Fuzzy-AHP technique. Priority wise categorization of usable-security factors helps practitioners for the motivation on fulfilling the user's demand and enhancement of the level of security for longer duration. This work contributed toward the formation of a hierarchy, which is valuable in designing usable-security [7]. With the help of the contribution, security developers shall be able to identify the vital usable-security factors, which further certifies the effective development of usable and secure web based healthcare management system design. This may facilitate practitioners to deliberate on the very significant usable-security factors first and to complete high satisfaction among customers with optimal maintenance.

TABLE. IV. IMPACT OF USABLE-SECURITY FACTORS

	Weights	**Percentage**	**Priority**
Confidentiality (C1)	0.104	10.40 %	5
Integrity (C2)	0.074	7.40 %	6
Availability (C3)	0.159	15.90 %	4
Effectiveness (C4)	0.185	18.50 %	3
Efficiency (C5)	0.237	23.70 %	2
Satisfaction (C6)	0.241	24.10 %	1

IV. DISCUSSION AND FINDINGS

There is increasing need for standardized levels of security in health care computing systems and networks. Also increasing this security should not badly influence the usability of web application. Hence usable-security is a big concern in modern era. In a healthcare environment, web based healthcare management system is becoming more complex, as its usage is regularly developing. This imposes essential to have a vastly secured web based healthcare management system. Security is one of the very noteworthy quality factors currently which is receiving maximum consideration of web based healthcare management system designers as well as users. This article has evaluated six usable-security factors while integrating usable-security during the web based healthcare management system development. Further, this paper provides an assistance to simply apply management plan during web based healthcare management system development. Major findings of the work are as follows:

- Address usable-security in order to improve secure life span of web based healthcare management system.
- Focusing on confidentiality, integrity, availability, effectiveness, efficiency, and satisfaction during web based healthcare management system development will improve usable-security.
- Satisfaction is the very noteworthy as well as suitable factor of usable-security to be considered to get secure service life of web based healthcare management system.

All in all, the results of this article prioritized the usable-security factors, which support the information that satisfaction should be taken at top priority when designing usable and secure web based healthcare management system.

V. Conclusion

In this research, identification of usable-security factors affecting the usability and security of web based healthcare management system has been done. Upon that, a hierarchical structure of factors is planned. Next, the opinion of 101 practitioners on the six usable-security factors has been taken. The practitioners are from web development industry as well as academic researchers. Using this opinion, weights of each factor has been calculated through Fuzzy AHP. It has been concluded that satisfaction is the very crucial factor between the six key usable-security factors. For the assurance of usable-security, practitioners shall initially focus on satisfaction for optimal maintenance of the web based healthcare management system.

References

[1] E. V. Bartlett, S. Simpson, Durability and reliability, alternative approaches to assessment of component performance over time, Available at:https://www.irbnet.de/daten/iconda/CIB8616.pdfAccessed on June 18 2019.

[2] Rajeev Kumar, Mohammad Zarour, Mamdouh Alenezi, Alka Agrawal, Khan R.A., (2019), Measuring Security-Durability through Fuzzy Based Decision Making Process, International Journal of Computational Intelligence Systems, June, 2019.

[3] Alka Agrawal, Mamdouh Alenezi, Suhel Ahmad Khan, Rajeev Kumar, Khan R.A., (2019), Multi-level Fuzzy System for Usable-Security Assessment, Journal of King Saud University - Computer and Information Sciences, April 2019.

[4] Kavita Sahu, R. K. Srivastava, Revisiting Software Reliability, Advances in Intelligent Systems and Computing, Springer, 2019.

[5] R. Kumar, S. A. Khan, R. A. Khan, (2016) Durability Challenges in Software Engineering, Crosstalk-The Journal of Defense Software Engineering, pp. 29-31.

[6] R. Kumar, S. A. Khan, R. A. Khan, (2015) Revisiting Software Security Risks, British Journal of Mathematics & Computer Science, Volume 11, Issue 6, 2015.

[7] Kavita Sahu, Rajshree, Rajeev Kumar, (2014) Risk Management Perspective in SDLC", International Journal of Advanced Research in Computer Science and Web based healthcare management system Engineering, pp. 1247-1251.

[8] R. Kumar, S. A. Khan, R. A. Khan, (2015) Durable Security in Software Development: Needs and Importance, CSI Communication, pp. 34-36, Oct 2015.

[9] Kavita Sahu, R. K. Srivastava, Soft Computing Approach for Prediction of Software Reliability, ICIC Express Letters-An International Journal of Research and Surveys, , pp. 1213-1222, 2018.

[10] Z. Zieliski, J. Chudzikiewicz, J. Furtak, An approach to integrating security and fault tolerance mechanisms into the military IOT, in: Chakraborty R., Mathew J., Vasilakos A. (Eds) secu. and faul. tol. in int. of things. (techn., comm. and com.). (2019), Springer.

[11] H. Assal, S. Chiasson, Think secure from the beginning, a survey with software developers, in: ACM, (2019), pp. 1-13.

[12] T. D. Oyetoyan, M. G. Jaatun, D. S. Cruzes, Measuring developers' software security skills, usage, and training needs, in: exploring security in software architecture and design, 1, (2019).

[13] J. Muñoz, F. Toutouh, Jaime, A review of dynamic verification of security and dependability properties, in: artificial intelligence and security challenges in emerging networks, 26, (2019).

[14] Y. Xu, X. Wen, W. Zhang, A two-stage consensus method for large-scale multi-attribute group decision making with an application to earthquake shelter selection, in: Com. & Indu. Eng., 116 (2018), pp. 113-129.

[15] X. Liu, Y. Xu, R. Montes, R-X Ding, F. Herrera, Alternative ranking-based clustering and reliability index-based consensus reaching process for hesitant fuzzy large scale group decision making, in: IEEE trans. on fuzzy sys. 27 (1) , 2019.

[16] L. Xia, Y. Xu, F. Herrera, Consensus model for large-scale group decision making based on fuzzy preference relation with self-confidence: Detecting and managing overconfidence behaviors, in: Inf. Fusion 52, (2019), pp. 245-256.

[17] K. Bylykbashi, D. Elmazi, K. Matsuo, M. L. Barolli, effect of security and trustworthiness for a fuzzy cluster management system in VANETs, in: cognitive systems research, 55, (2019), pp. 153-163.

[18] A.B. Saxena, M. Dawe, Trust framework for iaas—a tool based on security checks through standards and certifications. in: Satapathy S., Joshi A. (eds) inf.n and comm. Tech. for inte. Sys.. Smart Inno. Sys. and Techn., 107, (2019) Springer.

[19] C-W Chang, C-R Wu, and H-L Lin, Integrating fuzzy theory and hierarchy concepts to evaluate software quality, in: soft. qual. Jour. 16(2), (2008), pp. 263-276.

[20] P. R. Srivastava, A. P. Singh, K.V. Vageesh, Assessment of software quality: a fuzzy multi criteria approach, in: evolution of computation and optimization algorithms in software engineering: applications and techniques, IGI Global USA, 11, (2010), pp. 200-219.

[21] L. Mikhailov, Deriving priorities from fuzzy pairwise comparison judgements, in: fuzzy set. and sys., 134 (3), (2013), pp. 365-385.

[22] Y. Xu, F. J. Cabrerizo, E. Herrera-Viedma, A consensus model for hesitant fuzzy preference relations and its application in water allocation management, in: appl. soft compu., 2017, 58, pp. 265-284.

[23] R. Kumar, S. A. Khan, R. A. Khan, Revisiting software security: durability perspective, in: inter. jour. of hyb. info. tech. (SERSC) 8(2), (2015), pp. 311-322.

[24] Alenezi, M., Kumar, R., Agrawal, A., & Khan, R. A. (2019). Usable-security attribute evaluation using fuzzy analytic hierarchy process. ICIC Express Lett.-An Int. J. Res. Surv., 13(6).

[25] Alharbe, N., Atkins, A. S., (2014). A Study of the Application of Automatic Healthcare Tracking and Monitoring System in Saudi Arabia, International Journal of Pervasive Computing and Communications , Vol. 10, Issue 2, pp. 183-195.

Relationship Analysis on the Experience of Hospitalised Paediatric Cancer Patient in Malaysia using Text Analytics Approach

Zuraini Zainol[1], Puteri N.E. Nohuddin[2], Nadhirah Rasid[3], Hamidah Alias[4], A. Imran Nordin[5]
Department of Computer Science, Universiti Pertahanan Nasional Malaysia
Faculty of Science and Defence Technology, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia[1]
Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi 43600 Selangor[2, 3, 5]
Department of Paediatrics, Faculty of Medicine, Universiti Kebangsaan Malaysia, Cheras 56000 Kuala Lumpur[4]

***Abstract*—The purpose of this study is to analyse the keyword relationships of paediatric cancer patient's experiences whilst being hospitalised during the treatment session. This study collects data through 40 days of observations on 21 paediatric cancer patients. A combination of text analytical visualizations such as network analysis map and bubble graph to analyse the data is applied in this study. Through the analysis, keywords such as "cri" (crying), "lay" (laying), "sleep" (sleeping) and "watch" (watching) are found the most common activities that have been experienced by paediatric cancer patients when they were hospitalised. Based on this observation, it can be argued that these activities can be represented as the experience that they have whilst being in the hospital. Based on the findings, hospitalised paediatric cancer patient's experience is limited due to the treatment protocol that requires them to be attached to intravenous line. Therefore, most of their activities are focused in bed such as sleeping, playing with their mobile, watching video and so on in bed. This study also offers a novel approach of transforming cancer patient data into useful knowledge about keyword relationship in paediatric cancer patient's experience during their stay in the hospital. The incorporation of these two text analytics offers insights for researchers to understand the interesting hidden knowledge in the collection of unstructured data, and this information can be used by medical providers, psychologists, games designers and others to develop any applications that can assist their difficulties and ease their pain while warded in the hospital.**

Keywords—Patient experience; paediatric; keyword relationship analysis; bubble graph; text network analysis

I. Introduction

Millions of people have been diagnosed with cancer around the world affecting people regardless of age. Every year about 300,000 new cancer related cases are diagnosed in children and teens under the age of 19 every year [1]. Similarly, Ferlay et al. [2] report that 10.3% of Malaysians are at risk to die before the age of 75 years old because of cancer related illness and 10% of cancer sufferers are children [3]. Many factors could be the cause for these unfortunate children to have cancer for example genetic syndromes, low immune system and high doses of ionizing radiation are known cause of childhood cancer. Moreover, environmental factors including infection and immune response could be of risk factors [4, 5].

Today, the advancement of medical technology and the latest findings in drugs for cancer have contributed to the betterment of paediatric cancer management. The survival rates of child cancer patients are increasing but the diagnosis and the treatments provided can be extremely stressful and the treatment protocols planned for them can affect the life of a child and their family [3]. Available cancer treatments such as chemotherapy, surgery, radiotherapy, bone marrow transplantation and immune therapy amongst others, all of which may have an affect towards children's physical and psychological well-being [6, 7]. The physical and psychological affect could be caused by the medical and physical effects, psychological effects, and cognitive and neuropsychological effects.

Typically, paediatric cancer patients are required to be hospitalised during the treatment session. The unpleasant feeling to stay in the hospital is argued as one of the most frequently stated negative experience with the journey of paediatric cancer patients. Indeed, hospitalisation itself is a stressful situation, and a patient's stress level could increase significantly as the length of hospitalisation increases. Psychosocial effects could manifest as increased levels of anxiety, and concerns about dying which could lead to depression [8].

Based on the findings, hospitalised paediatric cancer patient's experience is limited due to the treatment protocol that requires them to be attached to intravenous line. Therefore, most of their activities are focused in bed such as sleeping, playing with their mobile, watching video and so on in bed. This study also offers a novel approach of transforming cancer patient data into useful knowledge through keyword relationship analysis in paediatric cancer patient's experience during their stay in the hospital. The incorporation of these two text analytics offers insights for researchers to understand the interesting hidden knowledge in the collection of unstructured data. Therefore, this information can be used by medical providers, psychologists, games designers and others to develop any programs or applications that can assist their difficulties and ease their pain while being warded in the hospital.

In this paper, the authors describe a work done to analyse the experience of hospitalised paediatric cancer patient in

Malaysia. The relationship analysis in the text analytics is used to analyse the qualitative data collected through ethnography study done at the paediatric oncology ward at Hospital Canselor Tuanku Mukhriz, Cheras, Kuala Lumpur, Malaysia. The results from the analysis would help to visualise the experience that paediatric cancer patients have whilst being warded in the hospital. The related work on paediatric cancer patients, patients experience in the hospital, relationship analysis and text analytics are described in Section II. The methodology of the study is presented in Section III. The results and the discussion of findings are presented in Section IV. Finally, Section V concludes this study.

II. Background and Related Works

Cancers in children are quite different from cancers affecting adults [9]. They tend to occur in different parts of the body, they look different under the microscope and respond differently to treatment. With the advancement of cancer treatments, many cancers can be cured. Doctors incline to practice the word 'remission' rather than the word 'cured'. Remission means there is no evidence of cancer following treatment. However, sadly, in some cases a cancer returns months or years later. In another research, it was reported that younger patients with cancer have greater fear of cancer recurrence (FCR) due to unpredictable disease outcome that leads to depression and anxiety [10]. Most pediatric cancer patients face with many kinds of experience, especially in the ward. Thus studying the experiences of pediatric patients are also important as authors are able to propose better services or care from human assistance like book reading and talking companion [11, 12] or technological assistance such as computers and games [13].

Patient experience is an increasingly important factor for improving services and designing products in many sectors like healthcare [14], computer gaming [13] and customer hospitality [15]. In this study, it focuses patients' experiences in healthcare treatments. Boote, Telford, and Cooper [16] pointed the prominence definitions of patient experience by defining terms 'disease' which is a physiologic and clinical abnormality and 'illness' which is defined as the subjective experience of the patients. While healthcare professionals can be considered the experts in disease, patients and healthcare consumers can be considered experts in illness which are the experiences of having the disease [17]. Boote et al. also argued that when both healthcare professionals and patients' perspectives are considered together, the analysis may offer a synergized knowledge which could offer new insights towards improving the healthcare industry. Studying patient experience has proven to be useful, in a number of areas. Initially, at the individual level in which private hospitals with patient-centered care has been championed for some years within the healthcare industry [18] and there is evidence that it can improve outcomes for patients, including patient safety, patient satisfaction and anxiety reduction [19]. Another area is that investigating patient experience can lead to a product design, for example computer games specifically designed and developed to assist patients endure the pains during their treatments [20, 21]. This paradigm of patient-physician experiences and interaction focuses on the inclusion of the patient in decisions about their care, and in relating the evidence about disease to their experience of illness in order to help make the most appropriate and also personalised treatment decisions for that individual.

Thus, a number of research used statistical method to analyse the patient experience data [14]. Nonetheless, data mining (DM) techniques were also commonly used in healthcare. In healthcare systems, massive data is available and data mining techniques are suitable to be used as a tool to analyze those mass data like classification such as rule set classifiers, decision tree algorithms, neural network architecture, neuro-fuzzy, and Bayesian network structure discovery techniques. DM is a well-known method to extract non trial knowledge from raw data in relation to strategic decision makings, prediction and insights. There are many techniques in DM such as Classification, Clustering, Association Rules, Bayesian Networks, and Decision Tree [22-25]. Galatas *et al* introduced a novel hybrid technique of feature selection and Naïve Bayes to determine patient satisfaction as significant indicators for decision makings and improving quality services of healthcare [14]. Another study was developed and tested a learning-based text-mining approach to facilitate analysis of patients' experiences of care and develop an explanatory model illustrating impact on Health-related quality of life (HRQOL) [26]. In hospitality domain, DM techniques are used to analyze customer behavior and predict the possible behavior of expected clients by using Classification, Regression, Link analysis, and Segmentation [14].

According to [27], Text Mining (TM) can be defined as an analytic process that is designed to explore the unstructured text documents in search of useful information and knowledge hidden from a large amount of text resources. TM has been widely applied in many application domains such as military knowledge [28-31], business analytics [32-34], documents analysis [35-37], social media issues [38-40], etc. In this research, TM which is one of DM method is explored as to find new knowledge through examining qualitative data of patient experiences while having chemo treatment [20, 21]. Through automation, TM approaches offer high throughput systems which can analyse much larger document collections than would be feasible to inspect manually. The advantages of this technique are two-fold: it is more likely to (i) detect non-trivial patterns, and (ii) lead to more statistically significant results. However, text interpretation requires both linguistic and domain knowledge, which make its automation a challenging task.

TM is corresponding to DM in its approach to knowledge extraction through identifying and analysing text patterns [41]. Where DM deals with data held within structured databases, TM approaches aim to find patterns in unstructured textual data. Unlike structured data, textual data and natural language require complex analysis in order to understand its content. Therefore, in textual data, normally context based keywords are communicated to the reader through the language used and the assumption of some background knowledge used of the users for interpretation. This context is difficult to ascertain in an automated process. Terms extraction and relationship analysis are used to extract knowledge from sets of unstructured text data. A research is conducted to investigate the relationships between verses and chapters at the keyword level in a Malay translated Tafseer. A combination technique of TM and network analysis is developed to discover non-trivial patterns and relationships of verses and chapters in the Tafseer. This is achieved through keyword extraction, keyword-chapter relationship discovery and keyword- chapter network analysis [42].

Another study reviewed substantial sets of cancer publications which involved expeditious growth of biomedical text [43]. This has lead the grown of TM techniques which are used to extract information about cancer diagnostics, treatment and prevention from the unstructured biomedical text and focused on presenting basic concepts of TM algorithms, tools and data sets [43]. These findings could theoretically assist many researchers to elect appropriate TM tools and datasets. Discussions on applying TM techniques to support cancer systems biology research are also been reviewed.

III. Framework of Document and Keyword Relationship Analysis

This study applies the enhancement of the proposed framework by Rasid et al. [21]. The optimised framework is applied to generate keyword relationships from the paediatric cancer patient observation forms which is recorded during their hospitalisation. The observation of paediatric cancer patients was conducted at a randomly chosen time-slot during the day; for example, in the morning (7 am to 3pm), afternoon (3pm to 11pm) and in the evening (11pm to 7 am). Fig. 1 shows the framework of document and keyword relationship analysis (FDKRA). The framework comprises of 2 main components: (i) Document Pre-processing and Keyword Extraction Analysis module and (ii) Keyword Relationship Visualisation module to present the experiment findings. Document and Keyword Extraction Analysis is the main analysis engine for extraction and ranking of documents and keywords / term relationships from the paediatric cancer dataset. Finally, the discovered document and keyword relationships are presented using two visualisation types including network graph and bubble graph.

In this study, the data was collected through the shadow observations method conducted in the paediatric oncology ward, 4D at Pusat Perubatan Universiti Kebangsaan Malaysia (PPUKM), Hospital Canselor Tuanku Muhriz, Cheras, Kuala Lumpur. The shadow observations were aim to collect paediatric cancer patient's activities/routines while they are hospitalised. All information gathered from the shadow observations are recorded using the observation form. The recorded data are then transcribed in the Excel format for data analyzing. The observation form is completely anonymous which only contains number of observation, patient's age, gender, days in ward, bed number, time, and treatment. The dataset underwent text pre-processing to ensure correct format for different terms and keywords before conducting text analysis.

The next element in the FDKRA describes the representation of the document, which refers to patients and keyword findings and analysis. Based on the keyword pattern found in the documents, the related patterns within a group are discovered. The relationship of the keywords should represent the content of patients' observation. This study is a part of the second component of FDKRA, focussing on grouping of selected keywords and visualising them as important keywords from experiences among paediatric cancer patients.

A. *Document Pre-Processing and Keyword Extraction Module*

This module consists of documents pre-processing and keyword extraction. The dataset that comprises of 21 plain text documents need to be prepared. In text analysis, the document pre-processing is the most important step as the raw dataset is often incomplete, inconsistent, contains many errors, etc. Poor data quality will affect the accuracy of TM results. The pre-processing helps to improve the quality of data and also, the accuracy and effectiveness of text analysis. This module is developed based on the concept of TF-IDF, counting and ranking the words in the given content, followed by selecting words that occur more than the threshold. This module generates a DKM-TFIDF, which is a m x n matrix that represents text documents (observation data) versus terms (frequent keywords). DKM-TFIDF tracks the term frequency for each term in all the observations. Thus, DKM-TFIDF can become a very large sparse matrix, depending on the number of documents and number of terms in each observation. DKM-TFIDF representation is a method to represent the documents as numeric structures. Representing text as a numerical structure is a common starting point for TM and analytics, such as search and ranking, creating taxonomies, categorisation, document similarity and text-based machine learning.

B. *Keyword Relationship Visualization Module*

The module consists of three different types of visualisations: (i) Word Cloud, a graphical representation of keyword frequency. Keywords are usually single words, with the importance of each keyword is differentiated with font size or colour; (ii) Text Network Analysis Graph illustrates relationships between survey documents and keywords. Keywords are displayed as round nodes and lines are used to represent the relationships between them; and (iii) Bubble Graph exhibits each document and its keywords. Each document is presented in a bubble and keywords associated with the documents are represented in sub-bubbles, following the DKM-TFIDF. Sub-bubbles are differentiated with colours and sizes.

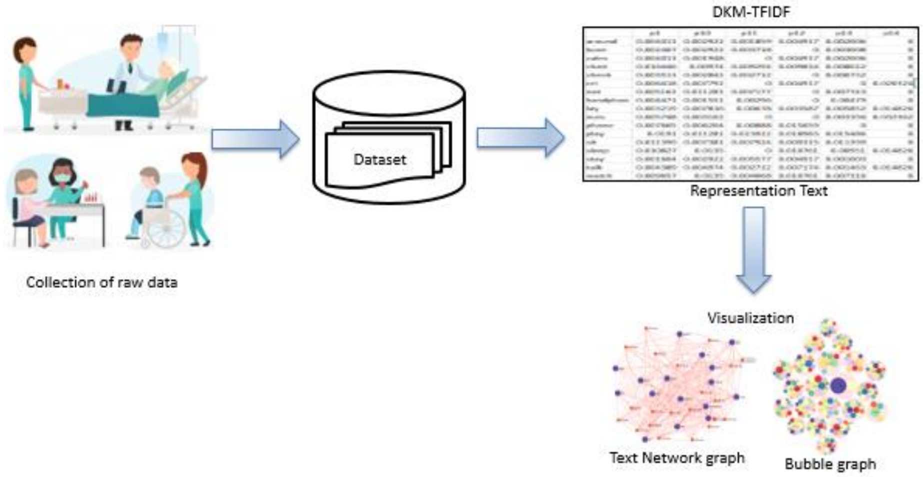


Fig. 1. Framework of Document and Keyword Relationship Analysis (FDKRA).

IV. Results and Discussion

This study applies the enhancement of the proposed. This section presents the experimental results using FDKRA. In this experiment, a set of 21 patients from the transcription of observation data is applied as the input. Each patient corresponds to a single document. The transcribed data is prepared as plain text documents. Text pre-processing is an important step in most text mining techniques and applications. It prepares the input data for consequent analysis. Low quality of text data affects the accuracy of TM results. Most text-based documents are often very noisy containing typos, errors and multiple acronyms for the same word. The pre-processing task can improve both the quality of data and accuracy, and effectiveness of text mining.

As shown in Fig. 2, the transcribed observation data is converted into a collection of text documents or corpus using the “tm” package in R. The text document is cleaned by removing numbers, symbols, punctuation marks, whitespace, etc., and converted all text into lowercase for standardisation. This is to ensure that multiple form of keywords such as “sleep” or, “SLEEP” are treated similarly in the experiment. After that, the words are tokenized by breaking up the text into discrete words. The next step is to remove all stop words (e.g., prepositions, pronouns, conjunctions, etc.) and reduced the existing words to their stems. This is to ensure that only the root of the word is presented in the document keyword matrix (DKM). In this study, some frequent keywords such as “mother”, “father” and “patient” are removed as these words are found common in the dataset. Besides that, the custom stop words including “get”, “can”, “im”, etc. are also removed from the dataset. These stop words are specific to the dataset that may not contain value to the dataset. In this experiment, the SnowballC package is applied for text document stemming. Such words “boring, “bored” and “bores” will be reduced to “bore” after stemming.

Most of TM tasks require data to be represented in the form of a matrix or vector – document term matrix (DTM) or term document matrix (TDM). The DTM or TDM describes the frequency of terms/keywords that occur in a collection of documents. As illustrated in Table I, the Document Keyword Matrix (DKM) consists of 475 terms (correspond to columns) extracted from 21 text documents (correspond to rows) with 84% sparsity. Sparsity refers to the threshold of relative document frequency for a term. The table below shows that 84% of the row entries in DKM contains zero entries. In other words, most keywords on DKM do not appear in most of text documents. For example, the keywords “*small*” (1), “*son*” (1), “*still*” (1), *“teach”* (1) and “*switch*” (1) are marked as zero in most text documents. Therefore, these less frequent keywords need to be removed.

Table I shows the results of removing sparse in DKM, with 38 keywords extracted from 21 documents with 31% sparsity which means that 31% of the entries contains zeroes (0). In other words, a 38 x 21 matrix is created representing 38 unique keywords and 21 text documents.

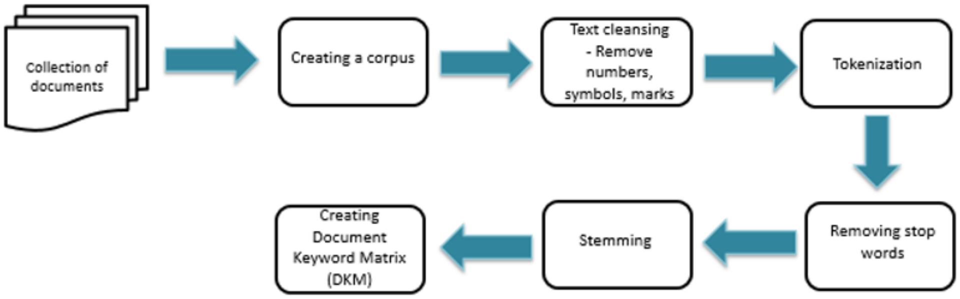

Collection of documents
Creating a corpus
Text cleansing - Remove numbers, symbols, marks
Tokenization
Removing stop words
Stemming
Creating Document Keyword Matrix (DKM)

Fig. 2. Overview of Text Pre-Processing Process.

TABLE. I. PARTIAL DKM AFTER REMOVING 31% SPARSITY

Docu-ments	Keywords								
	accompan	around	back	bed	Blanket	bore	bring	….	….	stand
p1	3	10	6	97	11	6	1	….	….	0
p2	0	1	0	6	2	2	0	….	….	1
p3	3	6	2	41	5	8	2	….	….	1
p4	4	7	2	53	4	11	3	….	….	0
….	….	….	….	….	….	….	….	….	…..	….
….	….	….	….	….	….	….	….	….	….	….
p21	1	0	0	2	0	0	0	….	….	1

Fig. 3 shows the visualisation of most frequently used keywords within the text documents, in the form of a word cloud. Word cloud is easy to understand and readable as it provides multiple choices of colours that symbolizing the keywords with different sizes. The size of keywords corresponds to the frequency of the terms. The larger font size corresponds to a higher frequency value. Based on our observation, the keywords "bed" (671), "sit" (348), "watch" (296), "sleep" (213), "play" (205), and "chair" (197) have the highest frequencies in the dataset.

After the text documents are pre-processed, the DKM is transformed into a TF-IDF representation, which highlights important terms in the survey. The term-weighting statistics is applied for identifying important keywords in a collection of text documents. Each keyword is assigned a weight, which represents its importance in the text document (Table II). The infrequent keywords with less weight are discarded. Thus, the list of terms in a text document can be arranged according to its importance. In this experiment, a subset of the most important terms is selected as keywords. The DKM-IDF also becomes an input for the network analysis map and bubble graph.

Table II shows the results of removing sparse in DKM-TFIDF, with 16 keywords extracted from 21 documents with 30% sparsity. In other words, a 16 x 21 matrix is created representing 16 unique keywords and 21 text documents. The new DKM-TFIDF becomes an input for the network analysis graph and bubble graph for keywords and patients.

Although the DKM-TFIDF consists of the summary of patients and their related keywords, it lacks of representing the visual data, particularly mapping the relationship between keyword and patients. Therefore, the DKM-TFIDF is transformed into a text network analysis. The text network analysis plots a text as a network graph where the nodes on the graph represents the specific keywords and patients (subjects). Fig. 4 illustrates the text network visualisation of the 21 patients with all the keywords. The blue nodes represent the 16 keywords such as "*around*", "*bore*", "*calm*", "*chair*", "*check*", "*cri*", "*eat*", "*handphon*", "*lay*", "*nur*", "*play*", *"sit", "sleep", "stay", "talk"* and "*watch*". On the other hand, the red nodes represent the 21 patients in the observation session. The light brown lines represent the linkages between patients and keywords. The thickness of colours for each connection is represented by the value of the occurrence of keywords in related documents, as per the DKM-TFIDF. For example, it can be clearly seen that the node (keyword) "*cri*" is linked to a group of nodes (patients) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 16, 17 and 20. These connections represent the node (keyword) "*cri*" is a common activity that had been experienced by almost all paediatric cancer patients. This relates to the observation that most of the paediatric cancer's patients are crying while receiving their chemotherapy treatments in ward.

Fig. 5 visualises a bubble graph that displays multiple bubbles of 21 patients with its important keywords developed from the DKM-TFIDF. This graph is developed using a Java Script. Each patient is represented in a pink bubble together with its related keywords. Each pink bubble (patient) may contain a group of multicolour sub-bubbles (selected important keywords). The size of sub-bubbles is dependent on the weight of a keyword (patient) and its associated important keywords. The legend on the left hand side contains 16 different coloured boxes indicating the type of keywords. The details of sub-bubbles can be further visualised by clicking on the selected bubble (patient).

For example, the bubble graph for patient 14 can be further visualised (see Fig. 6). The bubble that visualize patient 14 consists of five small multi-coloured sub-bubbles that relate to important keywords such as "nur" (nurse), "cri" (cry), "sleep", "talk" and "lay". A quick analysis of the bubble text visualization for keywords ("nurse"–"cry"–"sleep"–"talk"–"lay") shows that there is an important link between them. This relates to the observation that "The patient woke up from sleeping and crying and the nurse come to him. When his mother got back, he is just lying on the bed and talking to his mother".

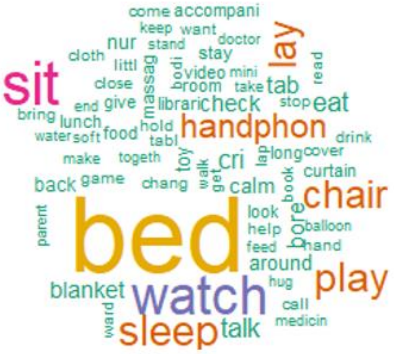


Fig. 3. Word cloud of the DKM with 100 Keywords.

TABLE II: PARTIAL DKM-TFIDF AFTER REMOVING 30% SPARSITY

Docu-ments	**Keywords**									
	around	**bore**	**calm**	**chair**	**Check**	**cri**	**eat**	**….**	**….**	**watch**
p1	0.004011	0.002407	0.004011	0.016446	0.003511	0.006418	0.005162	….	….	0.009657
p2	0.004555	0.009100	0.001948	0.004550	0.001820	0.013277	0.004550	….	….	0.000000
p3	0.005081	0.006775	0.000847	0.010162	0.002471	0.001694	0.007628	….	….	0.012973
p4	0.004894	0.007691	0.01398	0.011187	0.002040	0.002797	0.014397	….	….	0.014282
….	….	….	….	….	….	….	….	….	…..	….
….	….	….	….	….	….	….	….	….	….	….
p21	0.000000	0.000000	0.00000	0.000000	0.000000	0.00000	0.098079	….	….	0.00000

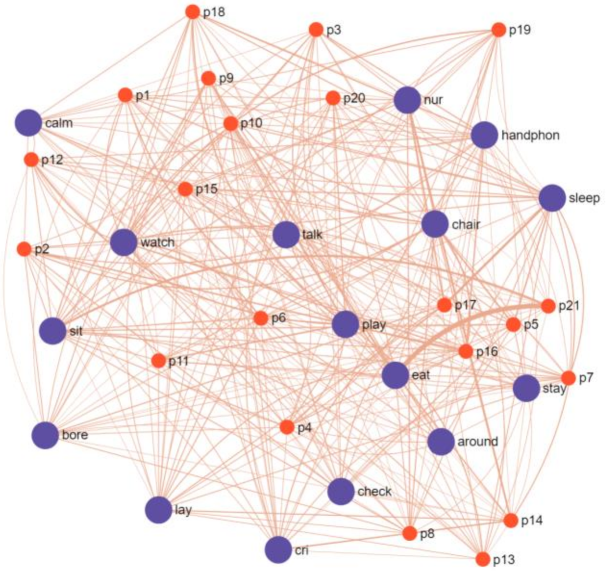


Fig. 4. DKM-TFIDF Text Network Analysis Map for Selected Important Keywords.

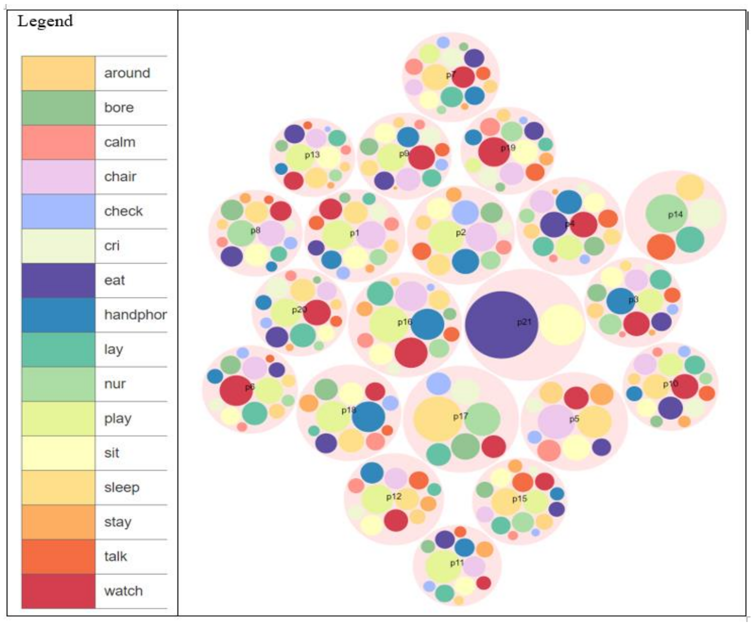


Fig. 5. Visualization of a Bubble Graph for 21 Patients and 16 Keywords.

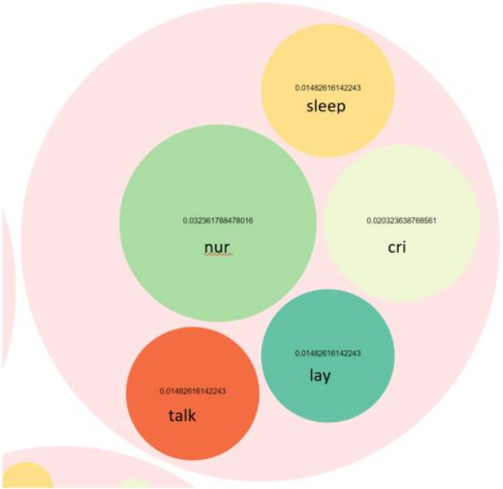


Fig. 6. Visualization of a Bubble Graph for Patient 14 with 5 main Keywords.

V. Conclusion

In this paper, the FDKRA framework was used to enable the collection and processing of the content of paediatric cancer patient's experience using two main components: (i) Document Pre-processing and Keyword Extraction Analysis module and (ii) Keyword Relationship Visualisation module to present the experiment findings. The results suggest that by using the keyword relationship visualisation module allows a clear presentation of unstructured data into a meaningful information. In our work, the collected observation data is hard to be interpreted given that the nature of the data is very unstructured. However, authors can present insight of the collected data to interpret patient's experience during their hospitalisation. The authors also argue that this strategy can be used for other qualitative study to allow researchers to have a quick understanding of the pattern of the collected data. This will ensure that the researchers would not miss any important pattern when they conduct the qualitative analysis.

Acknowledgment

The authors wish to thank the Universiti Pertahanan Nasional Malaysia (UPNM) and Pusat Keselamatan Siber (PKS), UPNM for funding and supporting this research. The authors also acknowledge the participation and cooperation received from patients and staffs from the paediatric oncology ward 4D, Hospital Canselor Tuanku Mukhriz (HCTM) Universiti Kebangsaan Malaysia and a little contribution from the FRGS Grant, Ministry of Education Malaysia (Grant No: FRGS/1/2016/ICT04/UKM/02/3) that has been contributed for the data collection phase.

References

[1] W. H. Organization, "International childhood cancer day: Much remains to be done to fight childhood cancer," Lyon, France: World Health Organization. Retrieved February, vol. 16, pp. 1-2, 2016.

[2] J. Ferlay, H. R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. J. I. j. o. c. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," vol. 127, no. 12, pp. 2893-2917, 2010.

[3] L. Penkman, L. Scott—lane, and W. J. J. o. p. o. Pelletier, "A psychosocial program for pediatric oncology patients: a pilot study of "The Beaded Journey"," vol. 24, no. 2, pp. 103-115, 2006.

[4] F. A. Cangerana Pereira, A. P. Mirra, M. d. R. Dias de Oliveira Latorre, and J. V. De Assunção, "Environmental Risk Factors and Acute Lymphoblastic Leukaemia in Childhood," Revista Ciencias de la Salud, vol. 15, no. 1, pp. 129-144, 2017.

[5] P. A. Buffler, M. L. Kwan, P. Reynolds, and K. Y. Urayama, "Environmental and genetic risk factors for childhood leukemia: appraising the evidence," Cancer investigation, vol. 23, no. 1, pp. 60-75, 2005.

[6] I. Hamzah, A. I. Nordin, H. Alias, N. Rasid, and H. J. A. S. L. Baharin, "Game Design Requirements Through Ethnography Amongst Pediatric Cancer Patients," vol. 24, no. 3, pp. 1567-1570, 2018.

[7] K. Enskär, L. J. N. C. von Essen, and Y. People, "Physical problems and psychosocial function in children with cancer," vol. 20, no. 3, 2008.

[8] I. Hamzah, A. I. Nordin, N. Rasid, and H. Alias, "Understanding Hospitalized Pediatric Cancer Patients' Activities for Digital Games Design Requirements," in Proc. 5th Int. Visual Inform. Conf., (IVIC) 2017: Springer, pp. 552-558.

[9] CHOP. "Pediatric Cancers Differ from Adult Cancers, and Need Different Treatment Plans, Say CHOP Experts." https://www.chop.edu/news/pediatric-cancers-differ-adult-cancers-and-need-different-treatment-plans-say-chop-experts (accessed 26 July 2019)

[10] J. V. Crist and E. A. Grunfeld, "Factors reported to influence fear of recurrence in cancer patients: a systematic review," Psycho-Oncology, vol. 22, no. 5, pp. 978-986, 2013.

[11] N. Salahieh. "Storytelling Can Be Therapeutic for Children Battling Cancer." https://uscstoryspace.com/2017-2018/salahieh/Fall_Midterm/ (accessed 19 July, 2019).

[12] "Chemo Companions." http://chemocompanions.org/ (accessed 26 July, 2019).

[13] M. Ghazisaeidi, R. Safdari, A. Goodini, M. Mirzaiee, and J. Farzi, "Digital games as an effective approach for cancer management: Opportunities and challenges," J. Edu. Health Promotion, vol. 6, 2017.

[14] G. Galatas, D. Zikos, and F. Makedon, "Application of data mining techniques to determine patient satisfaction," in Proc. 6th Int. Conf. on Pervasive Tech. Related to Assistive Environments, 2013: ACM, p. 41.

[15] I. Khan, R. J. Garg, and Z. Rahman, "Customer service experience in hotel operations: an empirical analysis," Procedia-Soc. and Behav. Sci., vol. 189, pp. 266-274, 2015.

[16] J. Boote, R. Telford, and C. Cooper, "Consumer involvement in health research: a review and research agenda," Health policy, vol. 61, no. 2, pp. 213-236, 2002.

[17] M. P. Pomey, D. P. Ghadiri, P. Karazivan, N. Fernandez, and N. Clavel, "Patients as partners: a qualitative study of patients' engagement in their health care," PloS one, vol. 10, no. 4, p. e0122499, 2015.

[18] I. Gabutti, D. Mascia, and A. Cicchetti, "Exploring "patient-centered" hospitals: a systematic review to understand change," BMC health services research, vol. 17, no. 1, p. 364, 2017.

[19] C. Doyle, L. Lennox, and D. Bell, "A systematic review of evidence on the links between patient experience and clinical safety and effectiveness," BMJ open, vol. 3, no. 1, p. e001570, 2013.

[20] N. Rasid, P. N. E. Nohuddin, H. Alias, I. Hamzah, and A. I. Nordin, "Using data mining strategy in qualitative research," in Int. Visual Inf. Conf. (IVIC), 2017, pp. 100-111.

[21] N. Rasid, P. N. E. Nohuddin, Z. Zainol, I. Hamzah, H. Alias, and A. I. Nordin, "Experience Mining Through Ethnography Study Among Pediatric Cancer Patients in Malaysia," Adv. Sci. Lett., vol. 24, no. 3, pp. 1562-1566, 2018.

[22] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[23] M. H. Dunham, Data mining: Introductory and advanced topics. Pearson Education India, 2006.

[24] S. Mourya and S. Gupta, Data mining and data warehousing. Alpha Science International, Ltd, 2012.

[25] P. N. E. Nohuddin, Z. Zainol, A. S. H. Lee, A. I. Nordin, and Z. Yusoff, "A Case Study in Knowledge Acquisition for Logistic Cargo Distribution Data Mining Framework," Int. J. of Adv. Appl. Sci., (IJAAS), vol. 5, no. 1, pp. 8-14, 2018.

[26] R. Wagland et al., "Development and testing of a text-mining approach to analyse patients’ comments on their experiences of colorectal cancer care," BMJ Qual Saf, vol. 25, no. 8, pp. 604-614, 2016.

[27] Z. Zainol, P. N. E. Nohuddin, M. T. H. Jaymes, and S. Marzukhi, "Discovering “interesting” Keyword Patterns in Hadith Chapter Documents," in Int. Conf. Information Commun. Tech. (ICICTM), pp. 104-108, 2016.

[28] Z. Zainol, P. N. E. Nohuddin, W. M. U. Noormanshah, and M. H. A. Hijazi, "Visualization of Context-Based Keyword Pattern Cluster Analysis on Tacit Knowledge Among Officer Cadets at Universiti Pertahanan Nasional Malaysia (UPNM)," Adv. Sci. Lett., vol. 24, no. 3, pp. 1550-1554, 2018.

[29] Z. Zainol, S. Marzukhi, P. N. E. Nohuddin, W. M. U. Noormaanshah, and O. Zakaria, "Document Clustering in Military Explicit Knowledge: A Study on Peacekeeping Documents," in Proc. 5th Int. Visual Inform. Conf. (IVIC), 2017, pp. 175-184.

[30] P. N. E. Nohuddin and Z. Zainol, "Discovering Explicit Knowledge using Text Mining Techniques for Peacekeeping Documents," Int. J. Business Inf. Sys. (IJBIS), 2020, in press.

[31] S. Marzukhi, N. H. Mohammad Daud, Z. Zainol, and O. Zakaria, "Framework of Knowledge-Based System for United Nations Peacekeeping Operations Using Data Mining Technique," in Proc. 4th Int. Conf. Information Retrieval Knowl. Man., (CAMP), pp. 18-22, 2018.

[32] A. S. H. Lee, Z. Yusoff, Z. Zainol, and V. Pillai, "Know your hotels well! -- An Online Review Analysis using Text Analytics," Int. J. Eng. & Tech., (IJET), vol. 7, no. 4.31, pp. 341-347, 2018.

[33] K. Berezina, A. Bilgihan, C. Cobanoglu, and F. Okumus, "Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews," J. Hospitality Marketing & Man., vol. 25, no. 1, pp. 1-24, 2016.

[34] N. F. Ibrahim, X. Wang, and H. Bourne, "Exploring the effect of user engagement in online brand communities: Evidence from Twitter," Comp. in Human Behavior, vol. 72, 2017.

[35] P. N. E. Nohuddin, Z. Zainol, F. C. Chao, M. T. James, and A. Nordin, "Keyword based Clustering Technique for Collections of Hadith Chapters," Int. J. Islamic Appl. Comp. Sci. Tech., (IJASAT), vol. 4, no. 3, pp. 11-18, 2015.

[36] W. M. U. Noormanshah, P. N. E. Nohuddin, and Z. Zainol, "Document Categorization Using Decision Tree: Preliminary Study," Int. J. Eng. & Tech., (IJET), vol. 7, no. 4.34, pp. 437-440, 2018.

[37] Z. Zainol, M. T. H. Jaymes, and P. N. E. Nohuddin, "VisualUrText: A Text Analytics Tool for Unstructured Textual Data," J. Phys.: Conf. Ser., vol. 1018, no. 1, p. 012011, 2018.

[38] Z. Zainol, S. Wani, P. N. E. Nohuddin, W. M. U. Noormanshah, and S. Marzukhi, "Association Analysis of Cyberbullying on Social Media using Apriori Algorithm," Int. J. Eng. & Tech., (IJET), vol. 7, no. 4.29, pp. 72-75, 2018.

[39] D. Mouheb, M. H. Abushamleh, M. H. Abushamleh, Z. Al Aghbari, and I. Kamel, "Real-Time Detection of Cyberbullying in Arabic Twitter Streams," in Proc. 10th IFIP Int. Conf. New Tech. Mobility Security (NTMS), 2019, IEEE, pp. 1-5.

[40] H. Mohamed, S. Marzukhi, Z. Zainol, T. M. T. Sembok, and O. Zakaria, "Semantic-based Social Media Threats Detection". Proc. 12th Int. Conf. Ubiquitous Information Man. and Comm., (IMCOM 18), , 2018.

[41] M. W. Berry and J. Kogan, "Text Mining. Applications and Theory," Wiley. 2010.

[42] S. Chua and P. Nohuddin, "Relationship Analysis of Keyword and Chapter in Malay-Translated Tafseer of Al-Quran," J. Telecommunication, Electronic and Comp. Eng. (JTEC), vol. 9, no. 2-10, pp. 185-189, 2017.

[43] F. Zhu et al., "Biomedical text mining and its applications in cancer research," J. Biomed. Informa., vol. 46, no. 2, pp. 200-211, 2013.

Prediction of Potential-Diabetic Obese-Patients using Machine Learning Techniques

Raghda Essam Ali[1], Hatem El-Kadi[2], Soha Safwat Labib[3], Yasmine Ibrahim Saad[4]
Teaching Assistant at Faculty of Computer Science, MSA University
Faculty of Computers and Artificial Intelligence, Giza, Egypt[1]
Associate Professor at Faculty of Computers and Artificial Intelligence, Giza, Egypt[2]
Associate Professor in Computer Science, Cairo University, Cairo, Egypt[3]
Associate Professor at Endemic Medicine, Hepatogastroenterology and Clinical Nutrition
Faculty of Medicine, Cairo University, Giza, Egypt[4]

***Abstract*—Diabetes is a disease that is chronic. Improper blood glucose control may cause serious complications in diabetic patients as heart and kidney disease, strokes, and blindness. Obesity is considered to be a massive risk factor of type 2 diabetes. Machine Learning has been applied to many medical health aspects. In this paper, two machine learning techniques were applied; Support Vector Machine (SVM) and Artificial Neural Network (ANN) to predict diabetes mellitus. The proposed techniques were applied on a real dataset from Al-Kasr Al-Aini Hospital in Giza, Egypt. The models were examined using four-fold cross validation. The results were conducted from two phases in which forecasting patients with fatty liver disease using Support Vector Machine in the first phase reached the highest accuracy of 95% when applied on 8 attributes. Then, Artificial Neural Network technique to predict diabetic patients were applied on the output of phase 1 and another different 8 attributes to predict non-diabetic, pre-diabetic and diabetic patients with accuracy of 86.6%.**

***Keywords*—Obesity; diabetes; nonalcoholic fatty liver disease; artificial neural network; support vector machine**

I. Introduction

Applying Machine Learning (ML) and Data Mining (DM) techniques in data mining studies are a main approach for using big quantities of accessible knowledge-based diabetes information. DM is one of the top priorities in science and medicine studies, this inevitably produces enormous quantities of data, due to the specific social impact of the effect of the severe disease. Consequently, without a doubt, for elements of clinical administration, diagnosis and management ML and DM techniques are of excellent interest. As a result, attempts were made to review the current literature on machine learning and approaches of data mining in diabetes research as part of this study.

Globally, obesity and diabetes have become huge public health problems, both associated multifactorial, complicated diseases [1]. However, many conditions can actually be avoided. Obesity is a notable increasing health issue; some call this the New World Syndrome [2]. It is described as an unusual or excessive accumulation of fat that poses a health danger.

Over 1.9 billion teenagers, 18 years of age and older, were overweight, more than 650 million of these were obese, in 2016 [3]. For non-communicable diseases as: cardiovascular illnesses, diabetes, musculoskeletal disorders, and types of cancers [4], it is a significant risk factor. Weight gain and body mass are essential to type 1 and type 2 diabetes development and increased incidence.

The definition of overweight and obesity is an extraordinary or excessive accumulation of fat that poses a health danger. The latest CDC (Center for Disease Control and Prevention) study demonstrates that the age-adjusted incidence of diagnosed diabetes increased dramatically from 3.5 to 6.6 for every 1000 population from 1980 and 2014 [5].

The Body Mass Index (BMI) [3] is a straightforward height to weight index usually used for adult's classification to be either underweight, overweight or obese. *'Overweight*' means a body mass index (BMI) of 25-29.9 kg/m² and *'Obese'* means a BMI of greater than 30 kg/m².

Overweight and obesity are powerful risk factors for type 2 diabetes and contribute significantly to precocious death. These metabolic disorders in the Eastern Mediterranean region are growing rapidly among adolescents. Adult data all over 16 countries in the Mediterranean region from age of 15 years and older show the highest levels of overweight and obesity such as in Egypt, Bahrain, Jordan, Kuwait, Saudi Arabia and the United Arab Emirates [5].

Diabetes mellitus is a one of the chronic diseases that is characterized by hyperglycemia. It can trigger a lot of complications [6]. As a result of the increasing mortality in the latest years, in 2040, the world's diabetic patients will reach 642 million [7], this means that there will be one adult per each ten adults suffers from diabetes in the future according to the WHO (World Health Organization) statistics.

This frightening estimate must undoubtedly be faced. Diabetes can cause chronic harm and abnormality or impairment in the function of a specified bodily organ or different tissues including eyes, heart, nerves, kidneys and blood vessels [8].

Diabetes is subdivided into two classifications, Type one Diabetes (T1D) and Type two Diabetes (T2D) [9]. T1D patients usually are younger, mainly under the age of 30 years. The common symptoms for these patients may include accelerated thirst, frequently urinated, high levels of glucose in blood [10]. This kind of diabetes cannot really be efficiently healed by using only oral drugs but also by using insulin treatment which considered to be very necessary. T2D which are usually linked with obesity, hypertension, fatty liver, dyslipidemia, arteriosclerotic and other diseases, is more frequently present in the mid-aged and elderly humans.

Diabetes can be diagnosed by evaluating glycated hemoglobin (HbA1c) taken from a blood sample. If the HbA1c reveals ⩾ 48 mmol/mol (6.5%) diagnosis may be alleged. HbA1c glycation is a measure of the plasma glucose concentration level and is used for both diagnosis and diabetes surveillance. HbA1c represents the mean plasma glucose of a patient in such a long period of time "*about three months*".

Nonalcoholic fatty liver disease (NAFLD) is frequently recorded in patients T2D [11], which has been proposed as a leading cause for NAFLD progression, or nonalcoholic steatohepatitis, probably reflect the quick succession of obesity and resistance to insulin in T2D.

Metabolic syndrome becomes more and more prevalent. It happens when there are a combination of a number of metabolic risk variables, such as obesity and insulin resistance. The risk of developing T2D increased by metabolic syndrome. Most usually, overweight individuals who are pre-diabetic or T2D produces much more insulin than nondiabetic individuals due to the greater bodily fat-muscle proportion. The possible explanation is that the body cannot use its insulin efficiently enough, which results in insulin resistance. It is therefore logical for the body to generate more insulin to offset. Furthermore, the growing quantity of insulin in the body progressively makes the body more resistant, it may also be seen as a comparable method of developing tolerances to drugs for drug users.

In order to diagnose an individual to be metabolism, Three out of Five requirements must be fulfilled: The elements of the Metabolic Syndrome, according to the WHO proposal [12] are: (1) *in abdominal obesity*: waist circumference of men ⩾ 102 cm and ⩾ 88 cm in women, (2) hypertriglyceridemia is greater than or equal to 150 mg/dl (1.695 mmol/L), (3) low HDL-C in men is less than 40 mg/dL (1.04 mmol/dL) and less than 50 mg/dL (1.30 mmol/dL) in women, (4) high blood pressure (BP) of greater than 130/85 mmHg and (5) high fasting glucose of more than 110 mg/dl (6.1 mmol/L).

Nonalcoholic Fatty Liver (NAFLD) could be classified as an added metabolic syndrome feature [13] with a certain resistance to hepatic insulin. The presence of fatty liver in patients with T2D and obesity has long been reported. Fatty liver has long been recorded in patients with type 2 diabetes and obesity [14]. It is generally regarded an incidental discovery, with little or no clinical significance. Sedentary lifestyle and bad nutritional habits contribute to weight gain and the chance of developing the metabolic syndrome and nonalcoholic fatty liver will increase eventually.

The importance of applying the proposed method for prediction of the diabetic patients who are already affected with both nonalcoholic fatty liver disease and obesity will help in minimizing the huge complications that results in an enormous health problems as an early diagnosis is the starting point for a successful living without the disease, it will also encourage and promote efficient interventions to monitor, prevent and manage diabetes mellitus disease and its complications in low and middle income nations, in particular.

The paper is organized as follows: Section 2 provides the required machine learning background understanding, Section 3 is divided into 3 subsections present the proposed system, Machine Learning techniques used and the methodological approach adopted, Section 4 provides the evaluation methods followed for valuation, Section 5 showed the proposed system results, Section 6 provides the conclusions.

II. Background

A. Machine Learning

The term "Machine Learning" is for many scientists identical to that of "Artificial Intelligence" [15], as the possibility of learning is the principal feature of an entity called intelligence. Machine learning is designed to build computer systems that can accommodate and benefit from their knowledge.

Knowledge discovery in database (KDD) [16] is an area that includes theories, methods and techniques, which attempt to create sense of information and derive helpful and valuable knowledge from it. The most significant stage in the KDD method is data mining, which idealize the implementation of machine learning algorithms in the analysis of data. It is regarded a multi-stepping process (selection, preprocessing, conversion, data mining interpretation and evaluation).

B. Machine Learning Types

The mining of data utilizes a variety of machine learning techniques to find hidden data patterns. These techniques are classified into three major categories: supervised learning techniques, semi-supervised and unsupervised learning techniques [16]. Physicians can use expert systems that are developed through machine learning techniques to help them easily diagnose and predict diseases given the importance of diseases diagnosis for humans, various studies on the classification methodologies have been conducted.

C. ML Applications on Medical Data

Medical diagnosis is an optimal field for algorithms of ML [17]. Many of them are recognized by patterns recognition on big quantities of data. To be effective in the field, an algorithm must be prepared, on comparatively few medical tests, to manage noisy and empty records of data.

Many studies were conducted in the area of machine learning in healthcare. Healthcare machine learning becomes one of the most researchers' priority. Insights can be obtained using different DM techniques and methods in hidden patterns recognition. These insights can also be used for forecasting of diseases and epidemics.

Kumar [18], showed that the goal of the different data mining techniques in health care systems is to highlight applications of data mining in healthcare depending on the nature of the dataset; as Artificial Neural Network and Support Vector Machine were applied in predicting

Parkinson’s disease with accuracy of 95%. And using statistical Neural Network in diagnosing breast cancer disease to improve the detection rate by 98.8%, and applied ANN, Multiple Association rule and immature bayed in predicting the heart disease.

Measuring the performance of DM techniques in healthcare prediction by applying multiple learning techniques (Basma Boukenze Hajar Mausannif & Abdelkrim Haqiq [19]); Decision tree, SVM and ANN, simulation of results showed that decision tree proved its performance in predicting chronic kidney failure disease than other learning techniques.

Also, when applying the same data mining techniques to specify the anemia type for anemic patients (M. Abdullah and S. Al-Asmari [20]), Decision Tree performs the best with accuracy result 93.75%. While using only SVM in classifying diabetes disease (Kumari and Chitra [21]), using Matlab 2010a tool to detect diabetes disease with accuracy of 78%.

Building decision tree and classification data mining methods help health care providers making better clinical decisions to identify chronic diabetes in early phases [22].

El-Halees and Shurrab [23], generated a model that can distinguish patients with normal blood disease from those who have blood tumor by using Multiple Association rules, classification techniques and ANN that resulted in accuracy of 79.45%.

III. Proposed System

A. Data Set

The dataset utilized were acquired form Al-Kasr Al-Aini, Faculty of Medicine, Cairo University. The dataset consist of 30 attributes, it were divided into two phases, the first phase consists of 8 attributes which are; Age, Sex, Schistosomiasis (Shisto), Alanine Aminotransferase (ALT), An aspartate aminotransferase (AST), Alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT) and nonalcoholic fatty Liver disease. The second phase consists of 8 attributes which are; Nonalcoholic Fatty Liver disease attribute (output of phase one), Weight, Height, Waist Circumference (WC), Fasting Blood Sugar (FBS), History of Hypertension, History of Diabetes and Hemoglobin A1C (HBA1C).

B. System Model

The suggested model in this paper comprises of two phases; the model starts from preprocessing step of filtering data then estimating the missing values, standardize data, normalize data after that handling the imbalanced data then verifying data to finally be ready for feature selection and extraction (Fig. 1).

Then, the first phase is developing Support Vector Machine algorithm to classify patients with nonalcoholic fatty liver disease (NAFLD). The learning algorithm was applied on 8 attributes and number of patients with NAFLD were detected and patients with other reasons of liver illness (alcohol, medication, etc.) were excluded to give off the results to either be 0; does not have liver disease or 1; affected with liver disease.

The second phase is developing a back propagation neural network that takes the output from phase 1 as an input in phase 2 in addition to another 7 attributes then train the artificial neural network algorithm with distinct topologies and range of epochs to achieve weights that give the optimum outcomes to categorize patients to three different classes; pre-diabetic, diabetic or nondiabetic patients “Fig. 2”.

1) Preprocessing: After reading the dataset file, preprocessing steps start in order to remove unwanted records that are presented in the dataset file by applying the following steps:

a) Filtering: by removing noise or unwanted data to display useful feature for prediction.

b) Estimating missing values: by calculating missing values as a weighted sum of linear interpolations from the closest accessible points. A total of five estimates from column-wise and five from row-wise, linear interpolation estimates for one-d are calculated. The best case was weighing; such that interpolation is equivalent to the average Lager 4-points of the nearest points in rows and columns (separated missing points far from the border).

c) Standardize and normalize: by rescaling attributes to the range of 0 to 1.

d) Handling imbalanced data: by adopting k-fold cross-validation steps in which data are randomly sorted, and then splited into k folds. after that, dividing the data using a common value of k=4 (four folds). The validation set consists of one fold, while the three remaining folds together are used for training. For each one of the validation sets, the validation accuracy is calculated and the final cross validation precision is averaged.

When you run 'k' cross validation rounds, one of the validation folds were used every round and the remaining folds were used for training. Its precision on the validation data was evaluated after being trained by the classifier. Average precision throughout the k round to obtain the final precision of cross-validation. Verify data by checking the dataset accuracy and inconsistency after data migration and proofreading data involving checking the data entered against the original document.

e) Export: release data to be ready for data mining.

*2) Feature selection and extraction:*There are number of characteristics for health information used for the system instruction. Noise, unauthorized or irrelevant information may also be present. The training dataset must be preprocessed in order to clean the data.

The Correlation Feature Selection (CFS) measure makes an evaluation for the subsets of features according to the following basics: “Good feature subsets contains features which are extremely associated with classification, but are not associated with each other” [24].

The primary goal of the feature selection method is to remove the redundancy and the non-relevant information. The result of improving classifier effectiveness and improving the accuracy is an increasing percentage of true positive predictive values. Decrease in accuracy is a result of nonrelevant features.

Feature Selection is considered to be one of the most important steps in the transformation phase of the KDD [16]. It is defined as the selection method of features from the area of study, which is more related and informative for model building. Feature selection [25] has many advantages that are relative to various elements of data analysis like improved data visualization and data realizing, reduced computational time and analysis time, and improved prediction precision.

3) *Machine learning techniques*

a) Support Vector Machine (SVM): Support Vector Machine (SVM) is considered to be one of the popular linear discrimination methods on the basis of a straightforward but strong powerful concept.

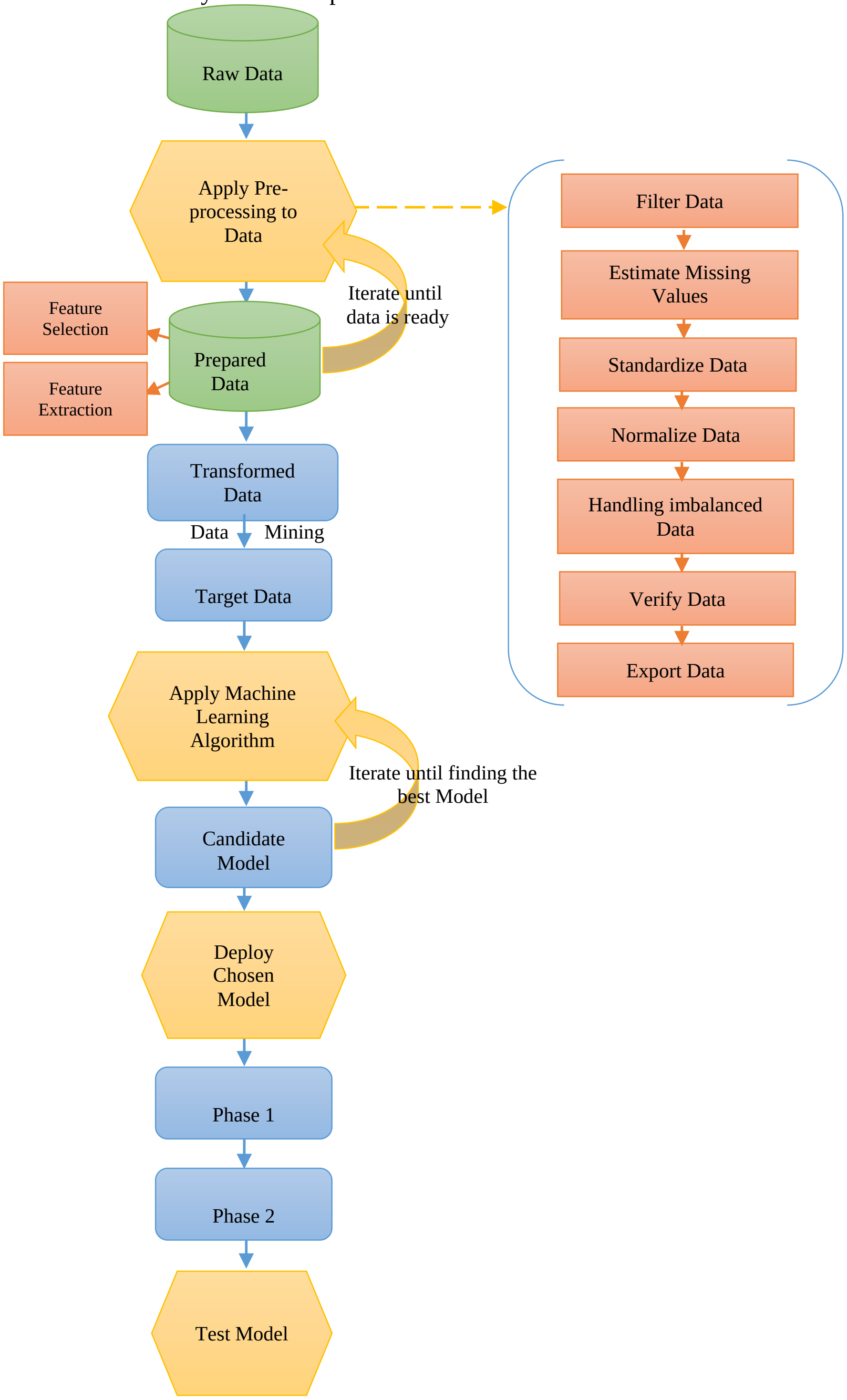


Fig. 1. Data Preprocessing.

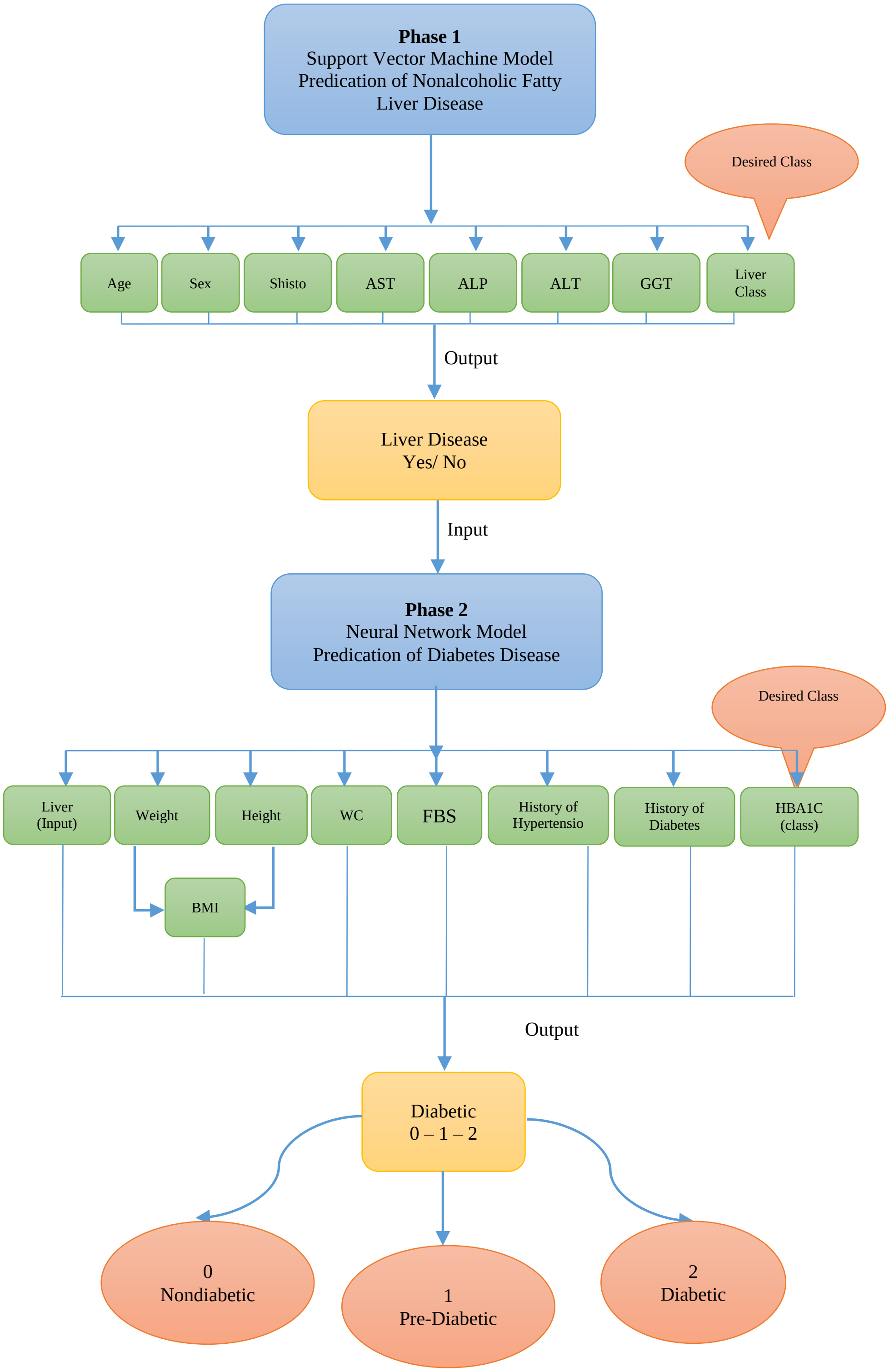


Fig. 2. Proposed System.

The first stage is to map the samples from the original entry to a high feature space so that the best way to separate samples is got. If its margin is largest then a hyperplane separating H is considered to be the top. The margin is the largest distance between two parallel hyperplanes to H on both sides that have no sample points between them [26, 27]. It comes from the concept of risk minimization (the expected loss assessment function as a miss-classification of samples) that the higher the margin, the higher the generalization error of the classifier.

Numerous kernels can be used in the Support Vector Machine models, including linear, polynomial, radial-based function (RBF) and Gaussian function:

$$K(X_i, X_j) = \begin{cases} X_i \, . X_j & Linear \\ (\gamma X_i \, . X_j + C)^d & Polynomial \\ exp(-\gamma \, |X_i - X_j|^2) & RBF \\ \tanh(\gamma X_i \, . X_j + C)^d & Sigmoid \end{cases} \tag{1}$$

Where $K(X_i, X_j) = \emptyset(X_i) \, . \emptyset(X_j)$ that is, the kernel function shows a dot product of input data that is mapped by transformation into the higher level feature space ϕ. Gamma is an adaptive parameter of some kernel functions.

Eventually, the RBF is the most common option in Support Vector Machines for Kernel types. This is primarily due to their localized and finite reactions throughout the true x-axis.

Algorithm 1 " Support Vector Machine (Phase 1)"

Detecting Nonalcoholic Fatty Liver Disease

```
Input: set of samples (input and output) for training pair samples;
the input samples are x1, x2…xn, and the output class is y.
    1. Finding Pair of samples in the training samples that are
closed
        candidateSV = {closest pair from classes that are
    opposite}
        do
        Find a violator sample
    2. Adding this sample to the Support Vector data samples
    Candidate_Vector = candidateSV_ Vector ∪ violator
    3. Pruning
        if any $\alpha_p < 0$ as a result of adding c to S then
        candidateSV = candidateSV / p
    4. Repeat till all points are pruned
         end if
```

b) Artificial Neural Network: Artificial Neural Network (ANN) method was used in the classification phase, as it utilizes complexity issues to be solved. The artificial neural network adapts itself by sequential training algorithm and its architecture and linked weights [24]. This paper utilized the learning algorithm for back propagation.

The Artificial Neural Network (ANN) is defined as a computational model made up of interconnected nodes that are called neurons arranged in layers; input, hidden and output. Each interconnection has a weight that changes during the training phase till adequate outcomes are achieved. ANN is used to model complex/nonlinear inter-relationships between inputs and outputs, for extracting significant patterns. In [26], ANN based classifier is used to model Diabetes dataset. The proposed ANN classifier has i- h - o configuration, where $i = 8$ (the number of attributes to the model inputs), h is the number of neurons in the hidden layer, where $h = 7$ (using one hidden layer), and o is the number of outputs that is equal one.

Algorithm 2 " Artificial Neural Network (Phase 2)"

Detecting Diabetes Disease

```
    1. Initialization step: set all weights equal to small random
    Values.
        Do while (from step 2 : 9)
    2. Iterate steps from 3 to 8: for each sample in the training set,
        Forward Phase:
    3. Each feature vector in the input are forward to the above
    layer (the hidden layer)
    4. Each hidden unit (Zj) sums its weighted i/p signals,
        $Z - i_{nj} = V_{aj} + \sum_{i=1}^{n} x_i v_{ij}$ , Where $V_{aj}$ is a bias
        Apply the transfer function
        $Z_j = 1/(1 + e^{-(Z - inj)})$
        send this value to all units that present in the above layer
    5. Compute the output:
        $Y - i_{nk} = W_{ok} + \sum_{i=1}^{n} Z_j w_{jk}$ , Where $W_{ok}$ is a bias
        $Y_k = 1/(1 + e^{-(Y - i_{nk})})$
        Backward Phase:
    6. Calculate the error in the output layer
        $\delta_{2k} = Y_k(1 - Y_k) * (T_k - Y_k), T_k$ is the target
    7. Computes its error information in hidden layers
        $\delta_{1j} = Z_j(1 - Z_j) * \sum_{k=1}^{m} \delta_{2k} w_{lk},$
        Update Phase:
    8. Update weights in all layers and bias
        $W_{jk}(new) = \eta * \delta_{2k} * Z_j + \alpha * W_{jk}(old)$
        $V_{ij}(new) = \eta * \delta_{1j} * x_i + \alpha * V_{ij} old$
    9. Test stopping condition.
```

IV. Evaluation Methods

A. Precision and Recall

Precision and recall are both common metrics for evaluating classifier efficiency and will be used widely in this dissertation. Precision is the proportion that when making a choice, the model properly predicts positive. To be more specific, precision is the number of positive instances properly identified divided by all number of positive examples “(2)”. Recall is the percentage of identified correct positive from all the current positives; it is the number of the correct positive classified exampled divided by the total number of true positive examples in the tested set.

Both high recall and precision are considered to be an ideal model. The F-measure “(5)” is the harmonic measure of precision and recall in a single measure [28]. The F- measure varies from 0 to 1, as a classified is a measure of 1 that completely captures precision and recall.

$$Precision \quad = \frac{TP}{TP+FP} \qquad (2)$$

$$Sensitivity \quad = \frac{TP}{TP+FN} \qquad (3)$$

$$Specificinty = \frac{TN}{TN+FP} \qquad (4)$$

$$F-measure = \frac{2(Precision)(Sensitivity)}{(Precision)+Sensitivity} \qquad (5)$$

Where

TN (True Negative): Negative case truly expected,

TP (True Positive): Positive case truly expected,

FN (False Negative): Negative case was positive but negatively expected,

FP (False Positive): Positive case was negative but positively expected.

B. Kappa Coefficient

Cohen's kappa statistics provide the second approach for datasets evaluation, which are 1 for an ideal classifier that usually classifies the right ones and 0 for a random classifier. The value of the kappa coefficient can be calculated using the following equation:

$$K = \frac{p_0-p_e}{1-p_e} \qquad (6)$$

Where, p_0 is the classification accuracy and p_e is the hypothetical accuracy of a random classifier on the same data.

V. Results

A. Support Vector Machine

Algorithm 1 " Support Vector Machine (Phase 1)"

Detecting Fatty Liver Disease

As shown in “Fig. 3”, it is obvious that the SVM with Gaussian and RBF kernel function give the best accuracy results. Thus, as both function return the same results RBF function was chosen to measure its precision and recall as shown in Table I.

B. Artificial Neural Network

Algorithm 2 "Artificial Neural Network (Phase 2)"

Detecting Diabetes Disease

The best performance was achieved using the primary dataset with overall 16 attributes, scaling each feature to a value between 0 and 1. Then, the classifier is trained as showed in Table II and demonstrates that the optimal accuracy results achieved in 50 iterations with 3 layers; 8 input nodes, 7 nodes in the hidden layer and 1 node in the output layer was 86.6% as shown in “Fig. 4”.

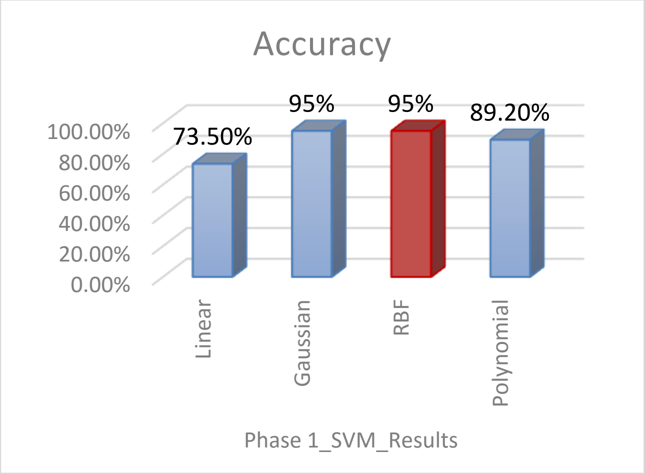


Fig. 3. SVM Accuracy Results.

TABLE. I. Phase 1_RBF Function Accuracy

RBF Function	
Measure	***Value***
Sensitivity	*0.90*
Specificity	*1.00*
Precision	*1.00*
Negative Predictive Value	*0.90*
False Positive	*0.00*
False Negative	*0.09*
Accuracy	*0.95*
F1 Score	*0.05*

TABLE. II. Phase 2_ANN Accuracy

Phase 2	**Artificial Neural Network**			
Iterations	*Input*	*Hidden*	*Output*	*Accuracy*
50	5	5	1	82.69%
100	5	5	1	82.69%
150	5	5	1	83.07%
200	5	5	1	82.30%
50	7	5	1	83.84%
100	7	5	1	84.61%
150	7	5	1	84.61%
200	7	5	1	84.61%
50	3	3	1	81.53%
100	3	3	1	82.30%
150	3	3	1	83.07%
200	3	3	1	83.07%
50	**8**	**7**	**1**	**86.56%**
100	8	7	1	85.38%
150	8	7	1	85.38%
200	8	7	1	85.38%

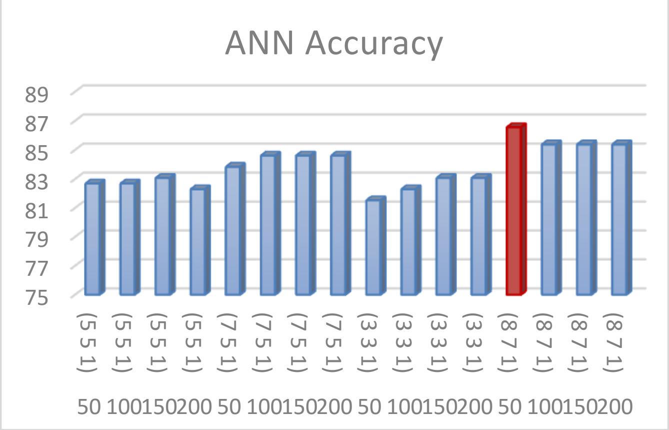


Fig. 4. ANN Accuracy Results.

In the experiments, when investigating the effect of the training data size on the classification accuracy, it has been noted that the size of the training set improves the classification accuracy. After analyzing the results, it can be concluded that the use of the hybrid system that combines SVM and ANN in one system is clearly preferable than using each classifier individually. Primarily, because SVM classifier is more efficient with binary class problem and very sensitive to the dimensionality of the feature vectors. In addition to the ANN algorithm which supposed to be a very flexible classifier. These combination leads to a powerful technique for classification problems.

VI. Conclusion and Future Work

Metabolic syndrome, non-alcoholic fatty liver and diabetes mellitus patients are at a growth of a very dangerous outcome like cirrhosis and morals for patients. In this study, a model for predicting chronic Diabetes mellitus was proposed.

The proposed model combines two machine learning techniques which are Support Vector Machine and Artificial Neural Network. The accuracy results showed that predicting nonalcoholic fatty liver disease by using the RBF kernel function in Support Vector Machine was 95% and by applying ANN classifier the findings obtained for optimal accuracy was 86.6% in 50 iterations with 3 layers; 8 input nodes, 7 nodes in the hidden layer and 1 node in the output layer.

Accordingly, good results on the obtained dataset showed that the proposed model performed out exemplary of the existing classifiers.

This analysis implies that patients with obesity, nonalcoholic fatty liver disease can lead to diabetes mellitus disease and more violent illnesses such as cirrhosis and mortality are expected to occur.

The results of this study exhibited the need of some further work to be done in the future. Firstly, more experiments will be required, since the imbalance of the dataset likely had a detrimental effect on the performance and the data processing was limited by the size of the dataset. Thus, more data should be involved to create a balanced dataset that would probably lead to a very important improvement in the performance for various learners, in order to make the research more universal. Secondly, more research is required to improve the quality of experimental information in preprocessing phase for data cleaning and estimation of the missing values. However, there are still many challenges in the medical research, and further work should be carried out to really advance this technology beyond laboratory demonstrations and disease prediction in order to restrict disease propagation.

Acknowledgment

I would like to express my great appreciation to Dr. Farid A. Mousa for his assistance, my gratitude to my family and my best friends.

References

[1] Bhupathiraju, S.N. and F.B. Hu, Epidemiology of obesity and diabetes and their cardiovascular complications. Circulation research, 2016. 118(11): p. 1723-1735.

[2] Ng, M., et al., Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. The lancet, 2014. 384(9945): p. 766-781.

[3] Al-Goblan, A.S., M.A. Al-Alfi, and M.Z. Khan, Mechanism linking diabetes mellitus and obesity. Diabetes, metabolic syndrome and obesity: targets and therapy, 2014. 7: p. 587.

[4] Iyer, A., S. Jeyalatha, and R. Sumbaly, Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774, 2015.

[5] Control, C.f.D. and Prevention, National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014. Atlanta, GA: US Department of Health and Human Services, 2014. 2014.

[6] Cichosz, S.L., Predictive models in diabetes: Early prediction and detecting of type 2 diabetes and related complications. 2016, Aalborg Universitetsforlag.

[7] Zou, Q., et al., Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 2018. 9.

[8] Krasteva, A., et al., Oral cavity and systemic diseases—diabetes mellitus. Biotechnology & Biotechnological Equipment, 2011. 25(1): p. 2183-2186.

[9] Devi, M.R. and J.M. Shyla, Analysis of various data mining techniques to predict diabetes mellitus. International Journal of Applied Engineering Research, 2016. 11(1): p. 727-730.

[10] Iancu, I., M. Mota, and E. Iancu. Method for the analysing of blood glucose dynamics in diabetes mellitus patients. in 2008 IEEE International Conference on Automation, Quality and Testing, Robotics. 2008. IEEE.

[11] Mills, E.P., et al., Treating nonalcoholic fatty liver disease in patients with type 2 diabetes mellitus: a review of efficacy and safety. Therapeutic advances in endocrinology and metabolism, 2018. 9(1): p. 15-28.

[12] Isomaa, B., et al., Cardiovascular morbidity and mortality associated with the metabolic syndrome. Diabetes care, 2001. 24(4): p. 683-689.

[13] Marchesini, G., et al., Nonalcoholic fatty liver disease: a feature of the metabolic syndrome. Diabetes, 2001. 50(8): p. 1844-1850.

[14] Ballestri, S., et al., Nonalcoholic fatty liver disease is associated with an almost twofold increased risk of incident type 2 diabetes and metabolic syndrome. Evidence from a systematic review and meta‐analysis. Journal of gastroenterology and hepatology, 2016. 31(5): p. 936-944.

[15] Peterson, D.M., The mind's new labels?: Review of RA Wilson and FC Keil (Eds.), The MIT Encyclopedia of the Cognitive Sciences. 2001, Elsevier.

[16] Kavakiotis, I., et al., Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 2017. 15: p. 104-116.

[17] Nilashi, M., et al., An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering, 2017. 106: p. 212-223.

[18] Kumar, R.N. and M.A. Kumar, Medical Data Mining Techniques for Health Care Systems. International Journal of Engineering Science, 2016. 3498.

[19] Boukenze, B., H. Mousannif, and A. Haqiq, Performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease. Int. Journal of Database Managment systems, 2016. 8(30): p. 1-9.

[20] Abdullah, M. and S. Al-Asmari, Anemia types prediction based on data mining classification algorithms. Communication, Management and Information Technology–Sampaio de Alencar (Ed.), 2017.

[21] Kumari, V.A. and R. Chitra, Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 2013. 3(2): p. 1797-1801.

[22] Daghistani, T. and R. Alshammari, Diagnosis of diabetes by applying data mining classification techniques. International Journal of Advanced Computer Science and Applications (IJACSA), 2016. 7(7): p. 329-332.

[23] El-Halees, A.M. and A.H. Shurrab, Blood tumor prediction using data mining techniques. Blood tumor prediction using data mining techniques, 2017. 6.

[24] Wiley, M.T., Machine learning for diabetes decision support. 2011, Ohio University.

[25] Jahankhani, P., V. Kodogiannis, and K. Revett. EEG signal classification using wavelet feature extraction and neural networks. in IEEE John Vincent Atanasoff 2006 International Symposium on Modern Computing (JVA'06). 2006. IEEE.

[26] Eljil, K.A.A.S., Predicting Hypoglycemia in Diabetic Patients using Machine Learning Techniques. 2014.

[27] Hashmi, S.F., A Machine Learning Approach to Diagnosis of Parkinson’s Disease. 2013.

[28] Frutuoso, D.G., SMITH-Smart MonITor Health system. 2015.

Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors

Marwa Almasoud[1], Tomas E Ward[2]
Information System Department
College of Computer and Information Science
Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia[1]
Insight Center for Data Analytics, Dublin City University, Dublin, Ireland[2]

***Abstract*—Chronic kidney disease (CKD) is one of the most critical health problems due to its increasing prevalence. In this paper, we aim to test the ability of machine learning algorithms for the prediction of chronic kidney disease using the smallest subset of features. Several statistical tests have been done to remove redundant features such as the ANOVA test, the Pearson's correlation, and the Cramer's V test. Logistic regression, support vector machines, random forest, and gradient boosting algorithms have been trained and tested using 10-fold cross-validation. We achieve an accuracy of 99.1 according to F1-measure from Gradient Boosting classifier. Also, we found that hemoglobin has higher importance for both random forest and Gradient boosting in detecting CKD. Finally, our results are among the highest compared to previous studies but with less number of features reached so far. Hence, we can detect CKD at only $26.65 by performing three simple tests.**

***Keywords*—Chronic Kidney Disease (CKD); Random Forest (RF); Gradient Boosting (GB); Logistic Regression (LR); Support Vector Machines (SVM); Machine Learning (ML); prediction**

I. Introduction

Chronic kidney disease (CKD) is a significant public health problem worldwide, especially for low and medium-income countries. Chronic kidney disease (CKD) means that the kidney does not work as expected and cannot correctly filter blood. About 10% of the population worldwide suffers from (CKD), and millions die each year because they cannot get affordable treatment, with the number increasing in the elderly. According to the Global Burden Disease 2010 study conducted by the International Society of Nephrology, chronic kidney disease (CKD) has been raised as an important cause of mortality worldwide with the number of deaths increasing by 82.3% in the last two decades [1, 2]. Also, the number of patients reaching end-stage renal disease (ESRD) is increasing, which requires kidney transplantation or dialysis to save patients' lives [1, 3, 4].

CKD, in its early stages, has no symptoms; testing may be the only way to find out if the patient has kidney disease**.** Early detection of CKD in its initial stages can help the patient get effective treatment and then prohibit the progression to ESRD [1]. It is argued that every year, a person that has one of the CKD risk factors, such as a family history of kidney failure, hypertension, or diabetes, get checked. The sooner they know about having this disease, the sooner they can get treatment. To raise awareness and to encourage those who are most susceptible to the disease to perform the tests periodically, we hope that the disease can be detected with the least possible tests and at low cost. So, the objective of this research is to provide an effective model to predict the CKD by least number of predictors.

In this paper, Section II reviews various research works that target the diagnosis of CKD using different intelligent techniques. Section III presents the dataset source and description. Section IV presents the methodology used for the prediction, including the data preprocessing steps and the modeling stage. Section V shows the results of the experiment and discusses the performance of ML algorithms in detecting CKD. Finally, Section VI includes the conclusion and future work of this work.

II. Literature Review

A. Related Work

In recent years, few studies have been done on the classification or diagnosis of chronic kidney disease. In 2013, T. Di Noia et al. [5], presented a software tool that used the artificial neural network ANN to classify patient status, which is likely to lead to end-stage renal disease (ESRD). The classifiers were trained using the data collected at the University of Bari over a 38-year period, and the evaluation was done based on precision, recall, and F-measure. The presented software tool has been made available as both an Android mobile application and online web application.

Using data from Electronic Health Records (EHR) in 2014, H. S. Chase et al. [6] identified two groups of patients in stage 3: 117 progressor patients (eGFR declined >3 ml/min/1.73m^2/year) and 364 non-progressor patients (eGFR declined <1 ml/min/1.73m^2) .Where GFR is a glomerular filtration rate that commonly used to detect CKD. Based on initial lab data recorded, the authors used Naïve Bayes and Logistic Regression classifiers to develop a predictive model for progression from stage 3 to stage 4. They compared the metabolic complications between the two groups and found that phosphate values were significantly higher, but bicarbonate, hemoglobin, calcium, and albumin values were significantly lower in progressors compared to non-progressors, even if initial eGFR values were similar. Finally, they found that the probability of progression in patients classified as progressors was 81% (73% − 86%) and non-progressors was 17% (13% − 23%).

Later in 2016, K. A. Padmanaban and G. Parthiban [7] aimed in their work to detect chronic kidney disease for diabetic patients using machine learning methods. In their research, they used 600 clinical records collected from a leading Chennai-based diabetes research center. The authors have tested the dataset using the decision tree and Naïve Bayes methods for classification using the WEKA tool. They concluded that the decision tree algorithm outweighs the Naïve Bayes with an accuracy of 91%.

A. Salekin and J. Stankovic [8] evaluated three classifiers: random forest, K-nearest neighbors, and neural network to detect the CKD. They used a dataset with 400 patients form UCI with 24 attributes. By using the wrapper method, a feature reduction analysis has been performed to find the attributes that detect this disease with high accuracy. By considering: albumin, specific gravity, diabetes mellitus, hemoglobin, and hypertension as features, they can predict the CKD with .98 F1 and 0.11 RMSE.

In the study carried out by W. Gunarathne, K. Perera, and K. Kahandawaarachchi [9], Microsoft Azore has been used to predict the patient status of CKD. By considering 14 attributes out of 25, they compared four different algorithms, which were Multiclass Decision Forest, Multiclass Decision Jungle, Multiclass Decision Regression, and Multiclass Neural Network. After comparison, they found that Multiclass Decision Forest performed the best with 99.1% accuracy.

H. Polat, H. D. Mehr, and A. Cetin [10] in their research used SVM algorithm along with two feature selection methods: filter and wrapper to reduce the dimensionality of the CKD dataset with two different evaluations for each method. For the wrapper approach, the *ClassifierSubsetEval* with the Greedy Stepwise search engine and *WrapperSubsetEval* with the Best First search engine were used. For the Filter approach, *CfsSubsetEval* with the Greedy Stepwise search engine and *FilterSubsetEval* with the Best First search engine were used. However, the best accuracy was 98.5% with 13 features using *FilterSubsetEval* with the Best First search engine using the SVM algorithm without mentioning which features were used.

P. Yildirim [11] studied the effect of sampling algorithms in predicting chronic kidney disease. The experiment was done by comparing the effect of the three sampling algorithms: Resample, SMOTE, and Spread Sup Sample on the prediction by multilayer perceptron classification algorithm. The study showed that sampling algorithms could improve the classification algorithm performance, and the resample method has a higher accuracy among the sampling algorithms. On the other hand, Spread Sub Sample was better in terms of execution time.

A. J. Aljaaf et al. [12] examined in their study the ability of four machine learning (ML) models for early prediction of CKD, which were: support vector machine (SVM), classification and regression tree (CART), logistic regression (LR), and multilayer perceptron neural network (MLP). By using the CKD dataset from UCI and seven features out of 24, they compared the performance of these ML models. The results showed that the MLP model had the highest AUC and sensitivity. It was also noticeable that logistic regression almost had the same performance as MLP but with the advantage of the simplicity of the LR algorithm. Therefore, in our study, we can use the LR algorithm as a start or a benchmark and then use more complex algorithms.

Lastly in 2019, J. Xiao *et al.* [13] in their study established and compared nine ML models, including LR, Elastic Net, ridge regression lasso regression SVM, RF, XGBoost, k-nearest neighbor and neural network to predict the progression of CKD. They used available clinical features from 551 CKD follow-up patients. They conclude that linear models have the overall predictive power with an average AUC above 0.87 and precision above 0.8 and 0.8, respectively

B. Dataset Concern

The dataset used in this study is a small dataset with small imbalance issue as will be described in Dataset section. Therefore, there are some concerns related to this dataset, which are an overfitting or generalization problem, imbalance, and the noise of the data. P. Yang et al. [14] in their review concluded that ensemble technique has the advantage of alleviating the problem of small size data by incorporating and averaging over multiple classifiers to reduce the probability of overfitting. Also, Deng et al. [15] found in their prediction of protein-protein interaction sites that the ensemble method can handle the imbalance problem and improve the prediction performance. Another survey by M. Fatima and M. Pasha [16] found that SVM provided improved accuracy to predict heart disease with the advantage of overfitting and noise [17].

III. Dataset

The dataset that supports this research is based on CKD patients collected from Apollo Hospital, India in 2015 taken over a two-month period. The data is available in the University of California, Irvine (UCI) data repository named Chronic_Kidney_Disease DataSet [18]. These data consisting of 400 observations suffer from missing and noisy value. The data includes 250 records of patients with CKD and 150 records of persons without CKD. Therefore, the percentage of each class is 62.5% with CKD and 37.5% without CKD. The ages of these observations are varied from 2 to 90 years old. It can be seen from Table I that the CKD dataset has 24 features including 11 numeric features and 13 nominal features, and the 25th feature indicates the classification or state of CKD.

TABLE. I. DESCRIPTION OF CKD DATASET

Name	Description	Type: unit/ values
Age (age)	Patient's age	Numeric: years
Blood pressure (bp)	Blood pressure of the patient	Numeric: mm/Hg
Specific gravity (sg)	The ratio of the density of urine	Nominal: 1.005, 1.010, 1.015, 1.020,1.025
Albumin (al)	Albumin level in the blood	Nominal: 0,1,2,3,4,5
Sugar (su)	Sugar level of the patient	Nominal: 0,1,2,3,4,5
Red blood cells (rbc)	Patients' red blood cells count	Nominal: normal, abnormal
Pus cell (pc)	pus cell count of patient	Nominal: normal, abnormal
Pus cell clumps (pcc)	Presence of pus cell clumps in the blood	Nominal: present, not present
Bacteria (ba)	Presence of bacteria in the blood	Nominal: present, not present
Blood glucose (bgr)	blood glucose random count	Numeric: mgs/dl
Blood urea (bu)	blood urea level of the patient	Numeric: mgs/dl
Serum creatinine (sc)	serum creatinine level in the blood	Numeric: mgs/dl
Sodium (sod)	sodium level in the blood	Numeric: mEq/L
Potassium (pot)	potassium level in the blood	Numeric: mEq/L
Hemoglobin (hemo)	hemoglobin level in the blood	Numeric: gms
Packed cell volume (pcv)	packed cell volume in the blood	Numeric
White blood cell count (wc)	white blood cell count of the patient	Numeric: cells/cumm
Red blood cell count (rc)	red blood cell count of the patient	Numeric millions/cmm
Hypertension (htn)	Does the patient has hypertension on not	Nominal: yes, no
Diabetes mellitus (dm)	Does the patient has diabetes or not	Nominal: yes, no
Coronary artery disease (cad)	Does the patient has coronary artery disease or not	Nominal: yes, no
Appetite (appet)	Patient's appetite	Nominal: good, poor
Pedal Edema (pe)	Does patient has pedal edema or not	Nominal: yes, no
Anemia (ane)	Does patient has anemia or not	Nominal: yes, no
Class	Does the patient has kidney disease or not	Nominal: CKD, not CKD

IV. METHODOLOGY

A. Data Preprocessing

Today's real-world datasets are susceptible to missing, noisy, redundant, and inconsistent data, especially clinical datasets. Working with low-quality data leads to low-quality results. Therefore, the first step in every machine learning application is to explore the dataset and understand its characteristics in order to make it ready for the modeling stage. This process is commonly known as data pre-processing.

1) Outliers: Outliers are extreme values located far away from the feature central tendency. Invalid outliers occur due to data entry errors, which are referred to as a noise in the data [19]. Medical data cannot be treated as other data in dealing with outliers since these outliers could be legitimate (valid) or important. For this reason, each outlier detected in the CKD dataset is checked to know if it is realistic or not. In this study, the extreme data points that go beyond the acceptable range medically have been treated as missing data and then modified as will be described in the missing data section. Box plots have been used to detect outliers in the CKD dataset, as Fig. 1 shows, there are some outliers detected for blood glucose random that reached 500 mg/dl. However, as mentioned in [20], the highest blood glucose level recorded in 2008 for a surviving patient reached 2,656 mg/dl. So, these outliers are legitimate and we should not change them.

In contrast, for potassium and sodium, three extreme data points are unacceptable. The highest potassium level observed was 7.6 mEq/L [21]. This means that a potassium level with 39 and 47, as shown in Fig. 2 is impossible and usually due to a mistake. Similarly, with sodium, as Fig. 3 shows, one extreme data point was detected, which is 4.5. Normally, sodium level should be between 135 and 145 mEq/L, and if it is less than 135, then the patient suffers from hyponatremia [22]. For this reason, a value of 4.5 is unacceptable or impossible.

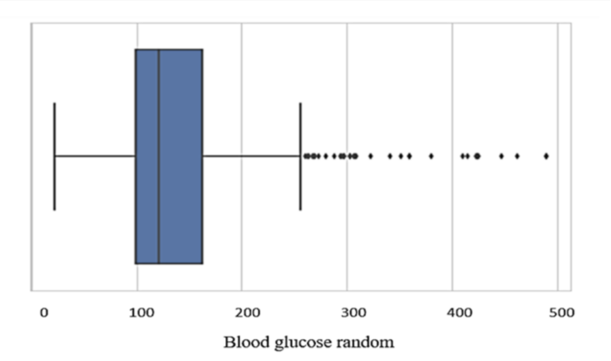


Fig. 1. Box Plot for Blood Glucose Random.

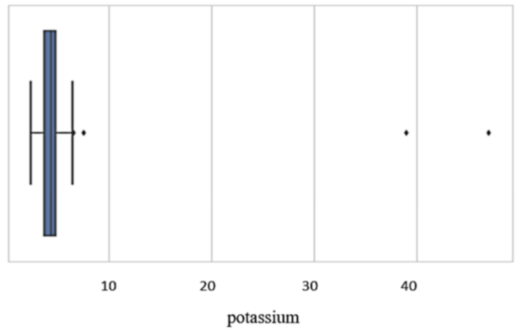


Fig. 2. Box Plot for Potassium.

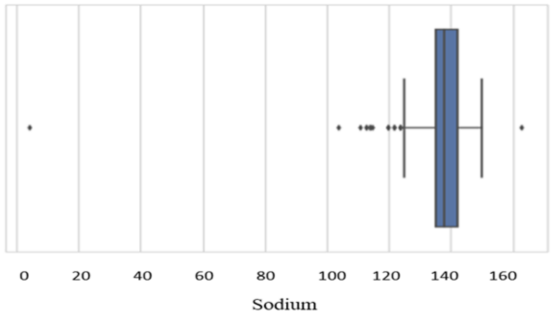


Fig. 3. Box Plot for Sodium.

2) *Missing Values:* In real-world datasets, missing data is a very common issue, especially in the medical area. Usually, every patient record and every attribute contains some missing values [23]. However, the chronic kidney disease dataset as shown in Fig. 4 has 96% of its variables having missing values; 60.75% (243) cases have at least one missing value, and 10% of all values are missing. There are different percentages of missing values for each variable, starting from 0.3% and reaching 38%, as shown in Table II.

Researchers in [9] used single imputation, such as mean and median, to impute the CKD dataset. However, according to Little's test [24], the missing values in CKD dataset are not missing completely at random (MCAR) with p-value <0.005. Therefore, single imputation cannot be used for handling missing values.

In this study, multiple imputations (MI) for replacing missing values in the CKD dataset. In multiple imputations (MI), missing values in the dataset are replaced *m* times, where *m* is usually a small number (from 3 to 10). We apply MI to produce five imputed datasets. The imputation process was based on linear regression for predicting continuous variables and logistic regression for categorical variables. Finally, we choose a dataset that has the nearest means and standard deviations for its variables to the original dataset.

3) *Data Reduction:* Data reduction means to reduce the number of features while maintaining a good analytical result. For this purpose, feature selection and features associations or correlation have been studied to remove redundant information.

a) Feature Associations: Pearson's correlation, Cramer's V, and ANOVA tests have been used to find relationships between variables. As shown in Fig. 5, and Fig. 6, there is a strong relationship between packed cell volume and hemoglobin and between hemoglobin and red cell count with the correlation coefficient of 0.89 and 0.79 respectively. Moreover, according to the ANOVA test, as shown in Table III, anemia also associated with PCV with p-value <0.001 $(2.16e^{-30})$. Another positive relationship was detected with a correlation coefficient of 0.68 between blood urea and serum creatinine.

TABLE. II. MISSING VALUES INFORMATION FOR EACH VARIABLE

Attribute Name	**Missing**		***Valid Number***
	Number	***Percent***	
Red blood cells	152	38.0%	248
Red blood cell count	131	32.8%	269
White blood cell count	106	26.5%	294
Potassium	90	22.5%	310
Sodium	88	22.0%	312
Packed cell volume	71	17.8%	329
Pus cell	65	16.3%	335
Hemoglobin	52	13.0%	348
Sugar	49	12.3%	351
Specific gravity	47	11.8%	353
Albumin	46	11.5%	354
Blood glucose	44	11.0%	356
Blood urea	19	4.8%	381
Serum creatinine	18	4.5%	382
Blood pressure	12	3.0%	388
Age	9	2.3%	391
Bacteria	4	1.0%	396
Pus cell clumps	4	1.0%	396
Coronary artery disease	2	0.5%	398
Diabetes mellitus	2	0.5%	398
Hypertension	2	0.5%	398
Anemia	1	0.3%	399
Pedal Edema	1	0.3%	399
Appetite	1	0.3%	399

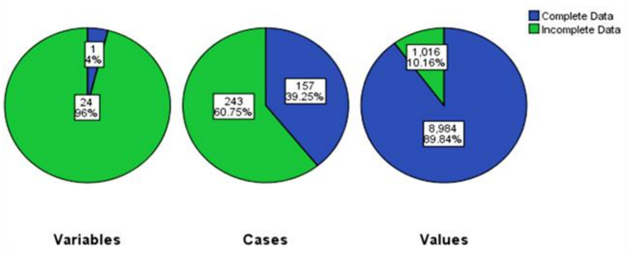


Fig. 4. Overall Summary of Missing Data in CKD Dataset.

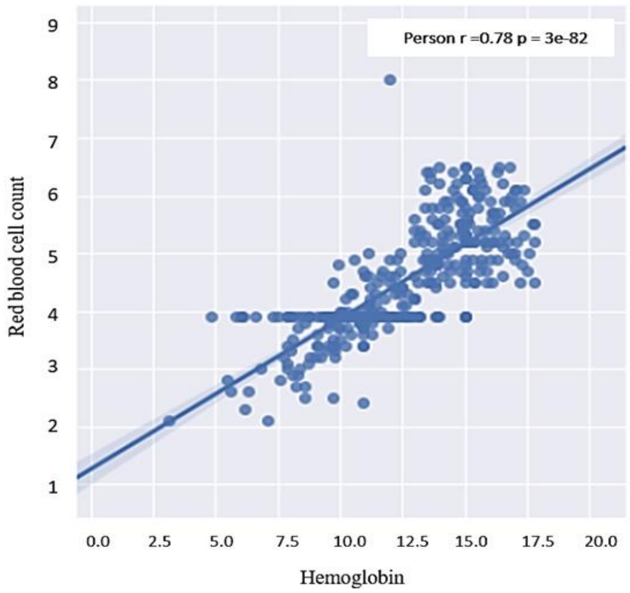


Fig. 5. Scatter Plot between Hemoglobin and Red Blood Cell Count.

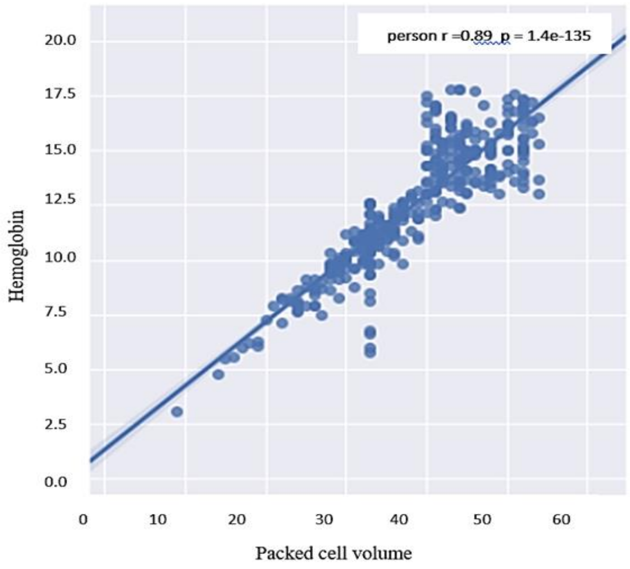


Fig. 6. Scatter Plot of PCV and Hemoglobin.

TABLE. III. ANOVA TEST RESULTS

Feature 1 Categorical	Feature 2 Numeric	P- value
Diabetes mellitus	Blood glucose random	$1.29e^{-26}$
Sugar level	Blood glucose random	$2.40e^{-49}$
Hypertension	Blood pressure	$6.13e^{-0.8}$
Anemia	Packed cell volume	$2.16e^{-30}$
Red blood cell	Red blood cell count	$2.36e^{-13}$
Pus cell	White blood cell count	0.00147

Since hemoglobin and serum creatinine have a stronger influence on the class attribute than their associated attributes, we decide to maintain them and remove the others as redundant attributes. Table IV shows the correlation between both numeric and nominal attribute and the class attribute.

TABLE. IV. CORRELATION BETWEEN ALL VARIABLES AND CLASS VARIABLE

	Numeric variable	Correlation coefficient		Nominal variable	Correlation coefficient
Pearson's correlation	Age	0.220	Cramer's V correlation	Albumin	0.730
	Blood pressure	0.296		Red blood cell	0.540
	Blood glucose random	0.399		Pus cell	0.420
	Blood urea	0.385		Pus cell clumps	0.214
	Serum creatinine	0.361		Bacteria	0.120
	Sodium	0.432		Hypertension	0.590
	Potassium	0.070		Diabetes Mellitus	0.544
	Haemoglobin	0.75		Coronary artery disease	0.236
	Packed cell volume	0.72		Appetite	0.393
	White blood cell count	0.222		Pedal edema	0.365
				Anemia	0.325
	Red blood cell count	0.666		Sugar level	0.432
				Specific gravity	0.687

Diabetes mellitus, sugar level, and blood glucose random almost measure the same thing, which is "sugar". The result proves the association by having p-value <0.001 ($1.29\, e^{-26}$) when testing the correlation between them, and 0.55% coefficient between sugar level and diabetes according to Cramer's V test. The same procedure was applied to other associated features. In the end, nine features have been removed as redundant features. These features are blood glucose random, blood pressure, packed cell volume, red blood cell count, red blood cell (nominal), anaemia, sugar level, pus cell, and blood urea.

b) Feature Selection: The process of selecting the most discriminating features in a given dataset is known as feature selection. This process is enhancing the model's performance, reducing overfitting, and reducing the cost of building a model. Filter feature selection methods [25] selects features that have a stronger relationship with the outcome variable independent to the learning model. Therefore, use a measure or test independent to the learning algorithm to assess a subset of features. In this study, mutual information measure has been used as a feature selection method. Mutual information [25] measures the dependence of any kind of relationships between random variables.

4) Data transformation: In data transformation, data is transformed into appropriate forms for mining purposes [26]. Data transformation includes normalization, which is the process of scaling the attributes' values to fall within a small specific range [26]. It is usually applied before feature selection and modeling stages because different scales of attributes complicate the comparison of attributes and influence the ability of algorithms to learn [23]. However, in this study min-max normalization has been applied on numeric data types. Another data transformation has been done on categorical variables. This is because some ML algorithms cannot handle categorical variables, especially in regression problems. Therefore, categorical variables with n values are dummied in LR and SVM classifiers by converting each of them into n-1 dummy variables [27].

B. Modeling

In the modeling stage, four machine learning algorithms have been applied to the dataset to assess their ability to detect CKD. These algorithms are logistic regression (LR), support vector machines (SVM), random forest (RF), and gradient boosting (GB).

1) Logistic regression: Logistic regression [28], also called logit model or logistic model, is a widely used model to analyze the relationship between multiple independent variables and one categorical dependent variable with the equation of the form:

$$\log\left[\frac{p}{1-p}\right] = a + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \qquad (1)$$

Where p is the probability of interest outcome, a is an intercept, β_1 ,……, β_i are β coefficients associated with each variable x , and $x_1 , \dots , x_i$ are the values of the predictor variables.

2) SVM: Support Vector Machines (SVM) [29] is a supervised learning model that is commonly used in classification problems. The idea of the SVM algorithm is to figure the optimal hyperplane that ideally separates all objects of one class from those objects of another class with the largest margins between these two classes. The objects that are far from the boundary are discarded from the calculation, while other data points that are located on the boundary will be maintained and determined as “support vectors” to get satisfactory computational efficiency [29]. The SVM algorithm has different kernel functions: radial basis function (RBF), linear, sigmoid, and polynomial. In this study, radial basis function has been chosen based on nested cross-validation results.

3) Ensemble method: Ensemble method [30] is a strategy for improving predictor or classifier accuracy. Ensemble method uses a combination of models to create an improved composite model to improve the performance. The main idea behind the ensemble technique is to group multiple “weak learners” to come up with a “strong learner”. Two popular techniques for constructing ensembles are bagging and boosting. Both boosting and bagging can be used for prediction as well as classification [26, 30]. Bagging is an ensemble technique where many independent predictors or learners are built and their results are combined using the majority vote, whereas in boosting, the predictors or learners are made sequentially not independently. This sequential method because each classifier “pays more attention” to the training tuples that were misclassified by the previous classifier through assigning weights for each of them [26]. Random forest algorithm is an example of the “bagging” technique, whereas the gradient boosting algorithm is an example of the “boosting” technique. Fig. 7 shows the bagging and boosting structure in selecting samples for training.

a) Random Forest: Random forest (RF) is a bagging ensemble approach proposed by Breiman [31] that based on a machine learning mechanism called “decision tree”. In a random forest, the “weak learners” in ensemble terms are decision trees [8, 32, 33]. Random forest imposes the diversity of each tree separately by selecting a random feature. After generating a large number of trees, they vote for the most common class. The random forest algorithm can deal with unbalanced data, it is robust against overfitting, and its runtimes are quite a bit faster [8, 31].

b) Gradient Boosting: Gradient boosting (GB) is an ensemble boosting technique that starts with “regression tree” as “weak learners”. In general, the GB model adds an additive model to minimize the loss function by using a stage-wise sampling strategy. The loss function measures the amount at which the expected value deviates from the real value. Stage-wise fashion put more emphasis on samples that are difficult to predict or misclassified. Unlike random forest, in GB, samples that are misclassified have a higher chance of being selected in training data [34]. GB reduces bias and variance and often provides higher accuracy, but the parameters should be tuned carefully to avoid overfitting. Therefore, nested cross-validation has been applied.

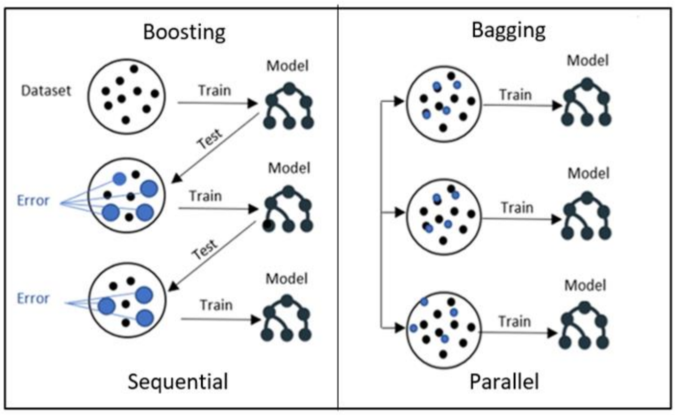


Fig. 7. Bagging and Boosting Structure.

V. Results and Discussion

The result of each classifier has been evaluated using different evaluation metrics and validated against overfitting using 10-fold cross-validation. The nested cross-validation approach also has been applied for the purpose of tuning the models’ parameters. The experiments are conducted using Python 3.3 programming language through the Jupyter Notebook web application. Several libraries from *Sciket-learn* [35] have been used, which is a free software for the machine learning library in Python. The evaluation measures considered in this study are accuracy using F1-measure, sensitivity, specificity, and area under the curve (AUC).

Each model generates different outputs depending on the different values of its parameters. By using nested cross-validation, the best performance for LR was with C=1000 and penalty=L2 with an accuracy of 98.9% using F1 measure. For the SVM model different values of "C", "gamma", and "kernel" have been tested. The best performance for SVM was with C=1 and gamma=3, and kernel = "RBF" (radial basis function) with an accuracy of 97.9% using F1 measure. For both RF and GB, the best results were with a number of trees=50 and Max_depth=2, with an accuracy of 98.0% in RF and 99.1% in GB using F1 measure.

The experimental results of each model in terms of accuracy, F1-measure, precision, sensitivity, specificity, AUC are listed in Table V whereas the training and testing accuracies based on 10-fold cross-validation are listed in Table VI.

From the evaluation results, as Fig. 8 shows, all models have an excellent performance against detecting CKD with an accuracy > 97% using hemoglobin, specific gravity, and albumin features. By focusing on specificity and sensitivity, it is seen that all models also have the same specificity of 99.3% except RF (96.6), which means that all models were accurate in identifying the negative or healthy subjects. On the other hand, the highest sensitivity was obtained using the RF algorithm at 99.6%, which represents the percentage of correctly identified CKD patients.

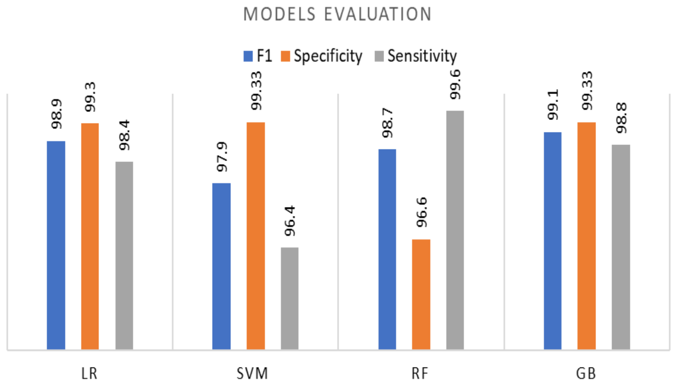


Fig. 8. Models Evaluation in Predicting CKD.

TABLE. V. THE PREDICTIVE PERFORMANCE OF ML MODELS

Classifier	**Accuracy**	**F1**	**Precision**	**Sensitivity**	**Specificity**	**AUC**
Logistic regression	98.75%	98.9 %	99.5 %	98.4 %	99.33 %	99.7 %
Support victor machines	97.5%	97.9 %	99.5%	96.4 %	99.33%	99.9 %
Random forest	98.5%	98.7%	98.0%	99.6%	96.6%	99.5%
Gradient boosting	99.0%	99.1 %	99.5%	98.8 %	99.33%	99.9%

TABLE. VI. THE TRAIN AND TEST ACCURACIES OF ML MODELS

	LR	**SVM**	**RF**	**GB**
Train	99.4 %	97.52%	98.5%	99.7%
Test	98.75 %	97.5%	98.5%	99.0%

Hence, we achieve the highest detection performance with the GB model. This performance is higher than the performance achieved by [12] using a multilayer perceptron algorithm (MLP), seven features, and single-point split with 98.4% F1- measure. Also, higher performance Compared to study [8], were 98.0% F1-measure have been achieved using RF and five features. According to study [8], which also estimated the cost of each of 24 tests in the CKD dataset, performing these three features for detecting CKD would cost only $26.65 while using all features will cost around $451.36.

Since the higher results were achieved using RF, and GB algorithm, we also investigate the importance of the features in each of them. As shown in Fig. 9, haemoglobin has the highest score, whereas Albumin has the lowest score in both RF and GB. Looking at RF, the degree of importance is convergent for all variables, approximately from 0.29 to 0.44. Whereas in GB, there is a significant difference between the degree of importance of haemoglobin (0.77) and other features. Then, according to our result, we conclude that haemoglobin has played an essential role in detecting CKD.

However, this research is subject to some limitations related to the dataset used. First, the size of the dataset is considered to be small (400 instances), which may influence the reliability of the results. Second, difficulty finding is another dataset that has the same features in order to compare the results of the datasets.

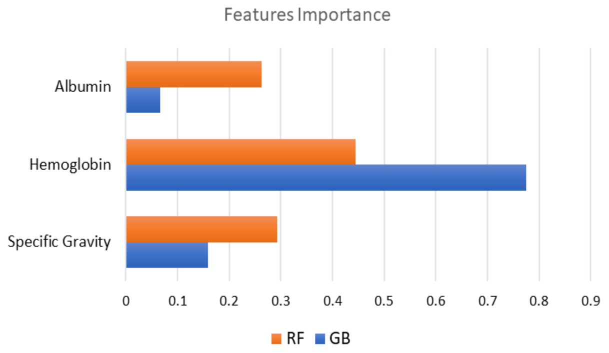


Fig. 9. Importance of Features in RF and GB Models.

VI. CONCLUSION AND FUTURE WORK

This work examines the ability to detect CKD using machine learning algorithms while considering the least number of tests or features. We approach this aim by applying four machine learning classifiers: logistic regression, SVM, random forest, and gradient boosting on a small dataset of 400 records. In order to reduce the number of features and remove redundancy, the association between variables have been studied. A filter feature selection method has been applied to the remaining attributes and found that there are haemoglobin, albumin, and specific gravity have the most impact to predict the CKD.

The classifiers have been trained, tested, and validated using 10-fold cross-validation. Higher performance was achieved with the gradient boosting algorithm by F1-measure (99.1 %), sensitivity (98.8%), and specificity (99.3%). This result is the highest among previous studies with less number of features and hence less cost. Therefore, we conclude that CKD can be detected with only three features. Also, we found that hemoglobin has the highest contribution in detecting CKD, whereas albumin has the lowest using RF and GB models.

Since the data used in this research is small, in the future, we aim to validate our results by using big dataset or compare the results using another dataset that contains the same features. Also, in order to help in reducing the prevalence of CKD, we plan to predict if a person with CKD risk factors such as diabetes, hypertension, and family history of kidney failure will have CKD in the future or not by using appropriate dataset.

ACKNOWLEDGMENT

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund. Special thanks to the deanship of scientific research at Princess Nourah bint Abdulrahman University for supporting this work.

REFERENCES

[1] J. Radhakrishnan et al, "Taming the chronic kidney disease epidemic: a global view of surveillance efforts," Kidney Int., vol. 86, (2), pp. 246-250, 2014.

[2] R. Lozano et al, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010," The Lancet, vol. 380, (9859), pp. 2095-2128, 2012.

[3] R. Ruiz-Arenas et al, "A Summary of Worldwide National Activities in Chronic Kidney Disease (CKD) Testing," Ejifcc, vol. 28, (4), pp. 302, 2017.

[4] Q. Zhang and D. Rothenbacher, "Prevalence of chronic kidney disease in population-based studies: systematic review," BMC Public Health, vol. 8, (1), pp. 117, 2008.

[5] T. Di Noia et al, "An end stage kidney disease predictor based on an artificial neural networks ensemble," Expert Syst. Appl., vol. 40, (11), pp. 4438-4445, 2013.

[6] H. S. Chase et al, "Presence of early CKD-related metabolic complications predict progression of stage 3 CKD: a case-controlled study," BMC Nephrology, vol. 15, (1), pp. 187, 2014.

[7] K. A. Padmanaban and G. Parthiban, "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease," Indian Journal of Science and Technology, vol. 9, (29), 2016.

[8] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in Healthcare Informatics (ICHI), 2016 IEEE International Conference On, 2016.

[9] W. Gunarathne, K. Perera and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)," in Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference On, 2017.

[10] H. Polat, H. D. Mehr and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," J. Med. Syst., vol. 41, (4), pp. 55, 2017.

[11] P. Yildirim, "Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction," in Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual, 2017.

[12] A. J. Aljaaf et al, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in 2018 IEEE Congress on Evolutionary Computation (CEC), 2018.

[13] J. Xiao et al, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of Translational Medicine, vol. 17, (1), pp. 119, 2019.

[14] P. Yang et al, "A review of ensemble methods in bioinformatics," Current Bioinformatics, vol. 5, (4), pp. 296-308, 2010.

[15] L. Deng et al, "Prediction of protein-protein interaction sites using an ensemble method," BMC Bioinformatics, vol. 10, (1), pp. 426, 2009.

[16] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9, (01), pp. 1, 2017.

[17] S. Karamizadeh et al, "Advantage and drawback of support vector machine functionality," in 2014 International Conference on Computer, Communications, and Control Technology (I4CT), 2014.

[18] L. Rubini. (2015). Chronic_Kidney_Disease DataSet, UCI Machine Learning Repository. Available: https://archive.ics.uci.edu/ml/datasets/ Chronic_Kidney_Disease.

[19] J. D. Kelleher, B. Mac Namee and A. D'arcy, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press, 2015.

[20] Michael and B. (2015). Highest blood sugar level. Available: http://www.guinnessworldrecords.com/world-records/highest-blood-sugar-level/.

[21] G. Gheno et al, "Variations of serum potassium level and risk of hyperkalemia in inpatients receiving low-molecular-weight heparin," Eur. J. Clin. Pharmacol., vol. 59, (5-6), pp. 373-377, 2003.

[22] D. A. Henry, "In The Clinic: Hyponatremia," Ann. Intern. Med., vol. 163, (3), pp. ITC1-ITC19, 2015. DOI: 10.7326/AITC201508040.

[23] A. J. M. Kaky, "Intelligent Systems Approach for Classification and Management of Patients with Headache," Liverpool John Moores University, 2017.

[24] R. J. Little, "A test of missing completely at random for multivariate data with missing values," Journal of the American Statistical Association, vol. 83, (404), pp. 1198-1202, 1988.

[25] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," Neural Computing and Applications, vol. 24, (1), pp. 175-186, 2014.

[26] J. Han, J. Pei and M. Kamber, Data Mining: Concepts and Techniques. Elsevier, 2011.

[27] M. A. Hardy, Regression with Dummy Variables. Sage, 1993(93).

[28] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," J. Clin. Epidemiol., vol. 49, (11), pp. 1225-1231, 1996.

[29] Z. Chen, X. Zhang and Z. Zhang, "Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models," Int. Urol. Nephrol., vol. 48, (12), pp. 2069-2075, 2016.

[30] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization," Mach. Learning, vol. 32, pp. 1-22, 1998.

[31] L. Breiman, "Random Forests," Mach. Learning, vol. 45, (1), pp. 5-32, 2001.

[32] T. G. Dietterich, "Ensemble methods in machine learning," in International Workshop on Multiple Classifier Systems, 2000.

[33] M. Khalilia, S. Chakraborty and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," BMC Medical Informatics and Decision Making, vol. 11, (1), pp. 51, 2011.

[34] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," Transportation Research Part C, vol. 58, pp. 308-324, 2015.

[35] F. Pedregosa et al, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, (Oct), pp. 2825-2830, 2011.

Hadoop MapReduce for Parallel Genetic Algorithm to Solve Traveling Salesman Problem

Entesar Alanzi[1], Hachemi Bennaceur[2]
Department of Computer Science
Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

***Abstract*—Achieving an optimal solution for NP-complete problems is a big challenge nowadays. The paper deals with the Traveling Salesman Problem (TSP) one of the most important combinatorial optimization problems in this class. We investigated the Parallel Genetic Algorithm to solve TSP. We proposed a general platform based on Hadoop MapReduce approach for implementing parallel genetic algorithms. Two versions of parallel genetic algorithms (PGA) are implemented, a Parallel Genetic Algorithm with Islands Model (IPGA) and a new model named an Elite Parallel Genetic Algorithm using MapReduce (EPGA) which improve the population diversity of the IPGA. The two PGAs and the sequential version of the algorithm (SGA) were compared in terms of quality of solutions, execution time, speedup and Hadoop overhead. The experimental study revealed that both PGA models outperform the SGA in terms of execution time, solution quality when the problem size is increased. The computational results show that the EPGA model outperforms the IPGA in term of solution quality with almost similar running time for all the considered datasets and clusters. Genetic Algorithms with MapReduce platform provide better performance for solving large-scale problems.**

***Keywords*—Genetic algorithms; parallel genetic algorithms; Hadoop MapReduce; island model; traveling salesman problem**

I. Introduction

Genetic algorithms (GAs) are stochastic search methods that have been successfully applied in many searches, optimization, and machine learning problems [1]. GAs are used to find approximate solutions in a reasonable time for combinatorial optimization problems. One of the main features of genetic algorithms is that they are inherently parallel. This makes them the most suitable for parallelization [2]. Parallel genetic algorithms (PGAs) can improve GAs to search in a huge solution space and reduce the total execution time. In general, there are three main models of parallel GAs: master-slave model, fine-grained model and coarse-grained also called island model.

The island model is a popular and effective parallel genetic algorithm because it does not only save time but also improves global research ability of GA [3]. Recently, the increasing volume of data requires high-performance parallel processing models for robust and speedy data analysis. Thus, the use of large-scale data-intensive applications has become one of the most important areas of computing.

Several technologies and approaches have been implemented to develop parallel algorithms. Hadoop MapReduce represents one of the most mature technologies. MapReduce programming model, proposed by Google [4], has become the prevalent model for processing a vast amount of data in parallel especially on a large cluster of computing nodes. Due to massive parallelization and scalability of MapReduce, it is used to develop parallel algorithms. It provides a ready-to-use distributed infrastructure that is scalable, reliable and fault-tolerant [4], [5]. The power of the MapReduce comes from the fact that it splits the data into smaller chunks processed in parallel by the mappers and merged by the reducers [6]. MapReduce aims to help programmers and developers to primarily focus on their applications on large distributed clusters, and hide the programming details of load balancing, network communication, and fault tolerance. Hadoop is the latest buzzword in cloud computing which implements the MapReduce framework. Hadoop is an Apache open-source software project designed for distributed parallel processing. It is designed to run applications on a big cluster of commodity nodes in a reliable, scalable and fault-tolerant manner. It is also designed to scale up from single servers to thousands of nodes. Each node in the cluster is a machine offers local computation and storage [7]. Hadoop deploys a master-slave architecture for computation and storage.

The basic design idea of MapReduce is inspired by two functions: Map and Reduce. Both Map Tasks and Reduce Tasks, which are written by the user, work on key/value pairs. A MapReduce application is executed in a parallel manner through two phases. In the first phase, all Map tasks can be executed independently. In the second phase, each Reducer task depends on the output generated by any number of Map task. Then, all Reducer tasks start executing their tasks independently [8]. The architecture of the MapReduce framework is shown in Fig. 1.

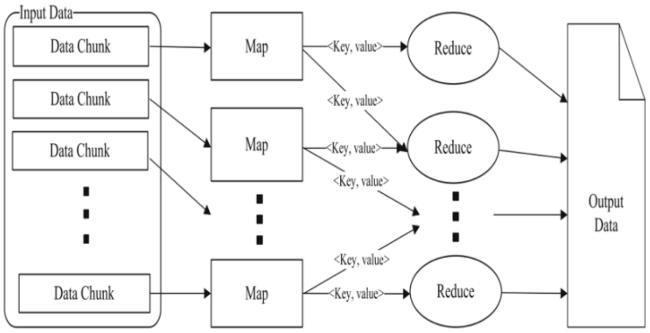


Fig. 1. The Architecture of the MapReduce Framework [8].

Traveling Salesman Problem (TSP) is one of the most common and combinatorial optimization problems in computer science and operations research. Given a set of cities and distances between them, the TSP goal is to find the shortest tour that visits all cities exactly once and returns to the starting city. TSP is easy-to-state but a difficult-to-solve problem since it is an NP-complete problem. TSP is considered enormously important because it can model a large number of real-world problems. Some of the applications include industrial robotics [9], job scheduling [10], computing wiring, DNA sequencing [11], [12], vehicle routing [13], and so forth.

Many methods have been developed to solve the TSP problem. These can be classified into two main categories; exact and heuristics. Exact methods guarantee to find the optimal solution of the problem whereas heuristic methods attempt to provide a good solution in a reasonable time [14]. Genetic Algorithms (GAs) are found to be one of the best metaheuristic algorithms for the TSP problems and yield approximate solutions within a reasonable time [15].

The parallel GAs using Hadoop MapReduce are not always guaranteed better performance than the sequential versions in term of execution time, that because of the overhead produced by the use of Hadoop MapReduce. One of the aims in this work is to understand if and when the parallel GA solutions show better performance.

For this work, two versions of parallel genetic algorithms are implemented using MapReduce to solve the TSP, a Parallel Genetic Algorithm with Island Model (IPGA) proposed in [16], and our proposed algorithm named Elite Parallel Genetic Algorithm (EPGA). We empirically assessed the performance of the two aforementioned PGA models with respect to a sequential GA on TSP problems, evaluating the quality of the solution, the execution time, the achieved speedup and the Hadoop overhead. The experiments were conducted by varying the problem size such as five TSP instances were exploited to differentiate the computation load. Additionally, varying population size and the cluster size were configured based on 4 and 8 parallel nodes. A total of 20 runs was executed for every single experiment of 3000 generations each.

The rest of the paper is organized as follows: Section 2 presents the related works. The third section presents the sequential genetic algorithms. Then, the proposed approach is described in Section 4. Afterward, we present the experiments, findings, and discussions in Section 5. Finally, Section 6 concludes the paper.

II. Related Works

In the last decade, there has been an increasing amount of literature on parallelizing genetic algorithms using MapReduce framework. The first work was an extension of MapReduce called MRPGA (MapReduce for Parallel GAs). Jin et al., [17] claimed that GAs cannot be directly expressed by MapReduce. They extended the original MapReduce by adding a second reduce phase at the end of each iteration to perform a global selection. The mapper nodes evaluate the fitness function. A local reducer for selecting the local optimum individuals and a second reducer produces the global optimum individuals as final results. Verma et al. [18] identified several shortcomings in the previous approach. Firstly, the mapper node performs the evaluation and the ReducReduce does the local and global selection, the bulk of the work—crossover, mutation and the convergence criteria are carried out by a single container. Hence, their approach decreased the scalability due to the sequential part of the coordinator. Secondly, mapper, reducer, and final reducer emitted “default key” with the value 1. In this respect, they changed the MapReduce model, and they did not apply any standards of the model whether grouping by keys or the shuffling.

Verma et al., [18] proposed a GA based on the traditional MapReduce model to solve ONEMAX problem. They considered one MapReduce job for each GA iteration. The Mapper nodes calculate the fitness values. Then, the Reducers implement selection and crossover operations. The default partitioner was overridden by a random partitioner in order to shuffle individuals randomly across different reducers to avoid overloading the Reducers. They confirmed that the GA can scale on multiple nodes with large population size. However, their model had a big IO footprint because the full population is saved to HDFS after each generation. Hence, big performance degradation was caused [16].

Huang and Lin [10] implemented a MapReduce framework to scale up the population for solving the Job Shop Scheduling Problem (JSSP) using GA. The authors used a large population size (up to 10^7) with fewer generations in order to reduce the overall MapReduce overhead for every generation. This study revealed that the GAs with larger populations were more likely to find good solutions as well as converge with fewer generations. Also, the effect of clusters size is presented, that show the speedup by increasing nodes in the cluster.

In [19], Subasi and Keco developed a Hadoop MapReduce model for parallelizing GA using one MapReduce phase for all generations of the genetic algorithm. Most of the processing was transferred from the reduce phase to map phase. This change reduced the amount of IO footprint because all processing data are kept in a local memory instead of HDFS. However, having a different population for each node leads to a species problem in the algorithm. To solve this problem, Enomoto et al., [20] applied migration strategy to improve population diversity in parallelization a GA using MapReduce, by exchanging individuals among subpopulations during the Shuffle phase. They utilized an ID for each island as a key to assign individuals to their sub-populations. Furthermore, they suggested a method to reduce unnecessary network IO in Shuffle tasks by reducing the number of individuals during migrations. This method eliminated half of the worst individuals in each sub-population after completing the map tasks (after GA- convergence). To maintain search efficiency, a number of individuals are recovered and created by applying a mutation operator at the beginning of each map phase. The results showed a significant improvement in the solution quality and execution time. In [21] the authors proposed a MapReduce hybrid genetic algorithm approach to solve the Time-Dependent Vehicle Routing Problem. The island model has been used for parallelizing the algorithm. The migration process has been carried out by changing the key (island ID) with a certain probability. They observed form the experiments that a large-scale problem with hundreds or thousands of nodes can be solved easily by adding more resources without any change in the algorithm implementation which is impossible in a single machine. Although this approach is interesting, it suffers from an overhead due to launching a MapReduce job for each GA iteration.

Apostol et al., [6] proposed new models for two well-known GA implementations, namely island and neighborhood models. They implemented the two models with two methods of handling sub-populations: island model with isolated subpopulation and neighborhood model with overlapping sub-population. The authors tested the algorithms on two optimization problems: the job shop scheduling problem and the traveling salesman problem with an instance of the problem of size 38 cities. The results showed that there was no significant difference between the two models in execution time, but the solution quality was higher for the neighborhood model over the island model. They proved a fact that the correct handling subpopulations formed by MapReduce can significantly improve the obtained results, but their work suffered from IO overhead between Mappers and reducers.

Ra et.al, [2] solved the TSP using GA on Hadoop MapReduce. They used multiple static populations with migration parallelization method. Iterative MapReduce jobs were used to implement Parallel GA. Each generation implemented a single MapReduce job. In the first job, the map tasks created an initial population and wrote them back to the HDFS. Then, the evaluation process started until a maximum number of generations, map tasks read populations from HDFS and sent them to reducers according to their Population Identifier. The reducers applied the GA operators, i.e. rank selection, greedy crossover and mutation with probability 2.1%, the population was evolved for a specific iteration number. All reducers wrote the best individuals of their populations and wrote the new population into HDFS. In order to share the best individuals, the best individuals were written with a different population identifier to migrate them to another sub-population in the next iteration. They measured the performance of their Parallel GA by comparing the sequential version of it (SGA) with the following algorithms-Sequential Constructive crossover, Edge Recombination crossover, and Generalized N-Point crossover. The results showed that the SGA came up with better solutions than other algorithms, but the SCX and SGA took almost the same time when the problem size increased. Moreover, they compared their SGA with MapReduce parallel GA, the MapReduce GA found better solutions. However, the SGA obtained solution faster than MapReduce GA for the small-sized problem because creation time for the map and reduce tasks impacts the solution time. However, when the problem size increased, SGA solution time increased, and MapReduce has almost the same run-time for all problem sizes.

Khalid et al.,[3] proposed a MapReduce framework implementation for GA with a large population using island parallelization technique. TSP was used as a case study to test the algorithm. A single MapReduce job was assigned to each generation. The Map task was responsible for fitness evaluation, crossover and mutation operations whereas the selection and re-population were done in reduce phase. The key represented the fitness of an individual while the value contained the individual itself. Accordingly, the intermediate pairs (key, value) were sorted and grouped according to fitness value. All individuals with the same fitness value were grouped into the same reducer. However, all copies of this individual were sent to an overloaded single reducer. When the GA converges, all the individuals were processed by that single reducer, so the parallelism would decrease as the GA converges and it would take more iterations. Furthermore, a single job was required for each generation which creates an overhead in term of execution time.

Rao and Hegde [22] proposed a novel method to solve TSP using the Sequential constructive crossover (SCX) on Hadoop MapReduce framework to deal with larger problem size. Iterative MapReduce was applied to specify a single job for each generation. The initial population was generated by the master node. Map tasks read the population from the HDFS and calculate the fitness function and then send them to reduce tasks with their population identifier. Afterward, the partitioner shuffled the individuals based on their sub-population identifiers. The remaining GA operators were performed in the Reduce phase. Upon the completion of iteration, all reducers wrote their best individuals and saved the new population into HDFS. An input file with 20 cities was used for the analysis purpose, and a hundred populations were initially created. They used a single-node Hadoop cluster on a single machine. The virtual machine and Cloudera open source Hadoop platform were used to deploy Hadoop. The results showed better performance after the various evolution of the genetic algorithm. However, using a single MapReduce job for each generation increased the overall overhead.

Ferrucci, Salza and Sarro [16] proposed a parallel genetic on Hadoop MapReduce platform based on three models, namely the global, grid and island models, they were used as a benchmark problem, the software engineering problem of configuring the Support Vector Machines (SVM) for inter-release fault prediction. They assessed the effectiveness of these models in terms of execution time, speedup, overhead and computational effort. The results revealed that the island model outperformed the use of Sequential GA and the PGAs based on the global and grid models. Furthermore, the overhead of the HDFS accesses, communication and latency impaired the parallel solutions based on global and grid models when executed on small problem instances. To speed up the execution of tasks, it was useful to reduce datastore operations as it happened with the island model where data store access was limited to the migration period only.

In this paper, we proposed a new MapReduce model to parallelize GAs named an Elite Parallel Genetic Algorithm using MapReduce (EPGA) in order to improve the population diversity. The Elite technique is inspired by the work of [23]. To the best of our knowledge, no literature proposes the Elite migrating based on Hadoop MapReduce to migrate the individuals between the master node and mapper/reducer node during the GA iterations.

III. Sequential Genetic Algorithms

The parallel adaptations are built on the base of the following SGA implementation, which is composed of a sequence of genetic operators repeated generation by generation, as described in Algorithm 1. The algorithm starts by generating randomly an initial population. It uses the tournament selection technique to select parents for the crossover operation. And it uses the inversion mutation operation which consists to swap randomly two nodes of the individual. The SURVIVAL selection is to select fitter individuals. After performing crossover operation survivor selection is used for selecting a next-generation population, as described in Algorithm 1.

ALGORITHM 1 SEQUENTIAL GENETIC ALGORITHM (SGA)

```
population ← INITIALIZATION (populationSize)
for i←1, MaxGenerationNumber do
for individual ∈ population
FITNESSEVALUATION (individual)
elite ← ELITISM (population)
    population ← population− elite
   parent1 ← TOURNAMENTSELECTION (population)
parent2 ← TOURNAMENTSELECTION (population)
child ← ESCX (parent1, parent2)
offspring ← offspring ∪ {child}
for individual ∈ offspring
INVERSIONMUTATION (individual)
for individual ∈ offspring
FITNESSEVALUATION (individual)
    population ← SURVIVALSELECTION (population,
offspring)
population ← population ∪ elite
```

IV. PARALLEL GENETIC ALGORITHMS USING MAPREDUCE

Island model is a popular and effective parallel genetic algorithm [3]. It reduces the communication overhead which is an eminent drawback in distributed computing and improves the global search ability of evolutionary algorithms [1]. When dealing with island models some aspects need to be considered:

- The migration interval: how often individuals are exchanged.
- The migration rate: the number of migrant individuals between sub-populations.
- The individual is chosen for migration.
- The individual replaced after the new individuals are received [6].

The following subsections describe in detail the algorithms used in this paper and how they are implemented with MapReduce.

A. Island Parallel Genetic Algorithm using MapReduce (IPGA)

The parallel genetic algorithm for the island model on MapReduce divides the population into several sub-populations. Each sub-population executed on a node called an island. Each island executes its sub-population a period of iterations independently from the other islands until a migration occurs Fig. 2. This period of consecutive generations before migration is defined as “migration period”. A MapReduce job is needed for each migration period. In this model, the numbers of Mappers and Reducers are coupled, each couple represents as an island. Each island has a specific identifier number.

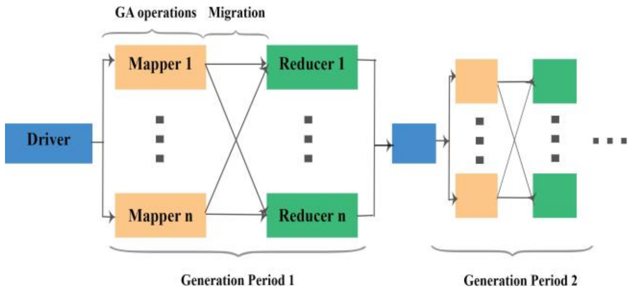


Fig. 2. The Flow of Hadoop MapReduce Implementation for IPGA.

The Mapper: As shown in the algorithm in Algorithm 2, mapper node is used to execute the generation periods, and at the end of the map phase, the migration function is applied. The Mappers output the subpopulation as records in the form (key, value). The key is distention island number, while the value contains the individual and its fitness value. Migration function selects best p % individuals (Migrant Individuals) from island i and migrates them to the next island (i+1) by changing their keys (island distention number).

ALGORITHM 2 MAP PHASE OF IPGA Map(key,value):

```
if population not initialized
    population ← INITIALIZATION(populationSize)
else
    Read population from HDFS
for i←1, GenerationPeriod
    for individual ∈ population
        FITNESSEVALUATION(individual)
    elite ← ELITISM(population)
    population ← population – elite
    parent1 ← TOURNAMENTSELECTION(population)
    parent2 ← TOURNAMENTSELECTION(population)
     child    ← ESCX (parent1, parent2)
    offspring ← offspring ∪ { child }
    for individual ∈ offspring
        INVERSIONMUTATION(individual)
    for individual ∈ offspring
        FITNESSEVALUATION(individual)
     population ← SURVIVALSELECTION(population, offspring)
     population ← population ∪ elite
end for
for i←1, MigrantIndividuals
       selectedIndividual ← GetBestIndividual (population)
       NextDestination ← islandNumber % totalNumberOfIslands
       remove worstIndividual from population
       EMIT  (selectedIndividual, NextDestination)
end for
for individual ∈ population
       EMIT (individual, islandNumber)
end for
```

The partitioner sends the individuals to the correspondent island (i.e., the reducer). While the reducer is used only to writes the sub-population received for its correspondent island into HDSF as shown in Algorithm 3.

ALGORITHM 3 REDUCE PHASE OF IPGA Reduce(key, values):

```
for individual ∈ population
    EMIT (individual, NullWritable.get())
```

B. *Elite Parallel Genetic Algorithm using MapReduce (EPGA)*

Migration Period is realized to propose MapReduce iterations in the previous subsection, but we must consider MapReduce overhead. MapReduce has processing overhead at the start and end times, overhead related to I/O to the data store (i.e., Hadoop Distributed File Systems (HDFS)), and communication overhead in Shuffle tasks. In IPGA, the data store access is limited to the migration phase only. In this study, we aim to reduce MapReduce jobs without decreasing the search efficiency. We propose to apply an Elite migration method in order to reduce the migration frequency without affecting the performance. In this model, the master node (Driver) will read all best individuals from each island at the end of each migration period, sort them by fitness order and share best %p individuals from the top of the list among all islands. The outline of the algorithm is as follows:

1) Each Mapper receives a sub-population to which it applies the GA from the HDFS.

2) Each Mapper performs the GA for a period of generations. An identification number associated with the island (island id) is assigned as a key. Then, a pair of the id and an individual is combined as a (key, value) pair respectively, and outputted to partitioner. If the current period of generations is not the first period, each Mapper reads the Elite individuals received from the master node and replaced them with the %p worst individuals. Elite individuals are added and executed within the current sub-population. Algorithm 4 shows the pseudo-code for the map phase.

3) Each partitioner receives the island id and individual given by the corresponding Map task. The partitioner assigns individuals to the Reducer by referring to the id of the island.

4) Each Reducer receives a subpopulation from its correspondent mapper, and selects best %p individuals, writes them into a separated file to HDFS. Also, outputs all other individuals to HDFS. Algorithm 5 shows the pseudo-code for the reduce phase.

5) If the maximum generation number not exceeded, the Master node reads the best individuals of all islands, sort them and selects best %p individuals and launches the next job.

6) If the maximum generation is achieved, then this process returns the global optimum individual and terminates. Otherwise, it repeats steps 1 to 5.

The flow of EPGA using MapReduce approach is shown in Fig. 3.

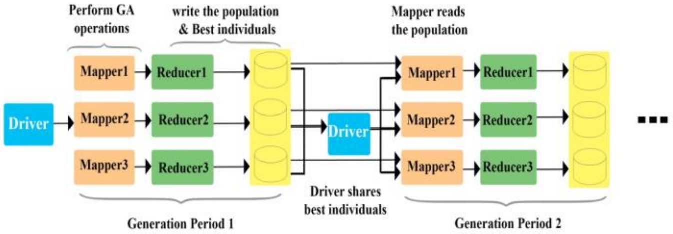


Fig. 3. The Flow of Hadoop MapReduce Implementation for EPGA.

ALGORITHM 4 MAP PHASE OF A GENERATION PERIOD OF EPGA
Map(key, value):

```
if population not initialized
    population ← INITIALIZATION(populationSize)
else
    Read population from HDFS
    Read EliteIndividuals list from Configuration
    Add EliteIndividuals to population
for i←1, MigrationPeriod
    for individual ∈ population
        FITNESSEVALUATION(individual)
    elite ← ELITISM(population)
    population ← population – elite
    parent1 ← TOURNAMENTSELECTION(population)
    parent2 ← TOURNAMENTSELECTION(population)
    child ← ESCX (parent1, parent2)
    offspring ← offspring ∪ { child }
    for individual ∈ offspring
        INVERSIONMUTATION(individual)
    for individual ∈ offspring
        FITNESSEVALUATION(individual)
    population ← SURVIVALSELECTION(population, offspring)
    population ← population ∪ elite
end for
for individual ∈ population
    EMIT (individual, islandNumber)
end for
```

ALGORITHM 5 REDUCE PHASE OF EPGA
Reduce(key, values):

```
sort population
select best individuals.
for individual ∈ population
    EMIT (individual, NullWritable.get())
write BestIndividuals to HDFS
```

V. Experiments and Results

The experimental evaluation of the proposed MapReduce algorithm is performed on the famous TSP problem. We conducted our experiments on the TSP Data sets provided by Andre Rohe [24]. The problems range in size from 131 cities up to 744,710 cities. Thus, we retained five datasets for a total 10 releases: ft70 (n=70, where n is the problem size)[25], xqf131 (n= 131), xqg237 (n= 237), bcl380 (n= 380), and rbu737 (n= 737). We chose these datasets because they are representing different degrees of the computational load for the fitness evaluation: small (ft70), medium (xqf131, xqg237, bcl381) and large (rub737).

We executed the two PGAs on two different cluster configurations (i.e., C4 and C8) characterized by a different number of nodes. For each PGA model (IPGA and EPGA), dataset (ft70, xqf131, xqg237, bcl380, and rbu737), population size (500, 1000, 2000, 5000 and 10000), and cluster configuration (4 nodes, and 8 nodes), we executed 20 runs. Thus, we executed a total of 2500 runs consisting of 2 × 5 × 5 × 2 × 20 = 2000 runs for PGAs and 5 × 5 × 20 = 500 for SGA.

The experiments are performed in an environment employing a private cloud platform of nine machines which compose the Hadoop cluster. We used a private Hadoop Cluster available at Computer Science department, Al-Imam Muhammad Ibn Saud Islamic University. All nodes have the same configuration to run a fair experiment as shown in Table I.

Table II summarized two different types of Hadoop clusters used in our experiments. SGA was executed on a single node, while for PGAs we exploited C4 and C8 clusters.

We employed the following settings for both SGA and PGAs:

1) Population varies in size 500, 1000, 2000, 5000 and 10000.

2) 3000 generations.

3) The elitism of 1 individual.

4) Tournament Selection for parent selection of size Enhanced Sequential Constructive crossover operator (ESCX) [14], with probability 1.

5) Inversion Mutation, with a probability of 0.5.

6) Survival Selection.

For the PGAs, we used the number of islands equal to the cluster size. We tuned the number of migrant individuals to best 10% of sub-population per island, and the migration period to 3 periods. The performance of the PGAs is measured with respect to execution time, solution quality, speedup, and overhead.

A. Execution Time

The execution time was measured in milliseconds (ms) using the system clock. We compared the computation time achieved by executing all generations of SGA and PGA. Fig. 4 shows the achieved execution times obtained over 20 runs for each dataset and with different population sizes (500, 1k, 2k, 5k, and 10k). We can observe that the PGAs (IPGA and EPGA) outperforms the SGA for the large datasets xqg237, bcl380 and rbu737 (Fig. 4(c), (d), and (e)), regardless of the number of parallel nodes used. And for the ft70 and xqf131 datasets, PGAs are better only when executed using more than 1k population (a and b, Fig. 4). This can be explained by the fact that, for small instance problems, the overhead due to communication between nodes is higher than the computational time. However, when the problem size or/and population size increases, the SGA execution time increases dramatically. We can observe from Fig. 4 that the execution time of the two PGA models using a C8 cluster, is better than using C4 cluster on all the datasets, so the use of more nodes allowed to further reduce the execution time. And the execution time of IPGA and EPGA is very similar time.

B. SpeedUp

The speedup is the ratio of the sequential execution time to the parallel execution time [16]. The speedup is calculated based on the following equation:

$$\text{Speedup} = \frac{\text{SGA time}}{\text{PGA time}} \qquad (1)$$

We compared the achieved speedup with respect to the ideal speedup. The ideal speedup is equal to the number of parallel nodes and corresponds to the situation when the SGA execution time is split among multiple nodes. Fig. 5 shows the speedup obtained by PGAs for all considered datasets. Both PGAs speed up the execution time with respect to SGA over all datasets of mean 7.2 × times by exploiting IPGA usingC8 cluster, 3.9 × times by exploiting IPGA using C4cluster. And 6.7 × times by exploiting EPGA using C8 cluster, 3.8× times by exploiting EPGA using C4 cluster. It is clear from the figure that, both PGAs tend to the ideal speedup value.

C. Solution Quality

The solution quality of the TSP problem was measured by calculating the error (Error%) of approximation of the best individual's fitness value and the TSPLIB optimum found on the website [24]. The error of the best path found with regard to the optimal tour in the TSPLIB is calculated as the given formula:

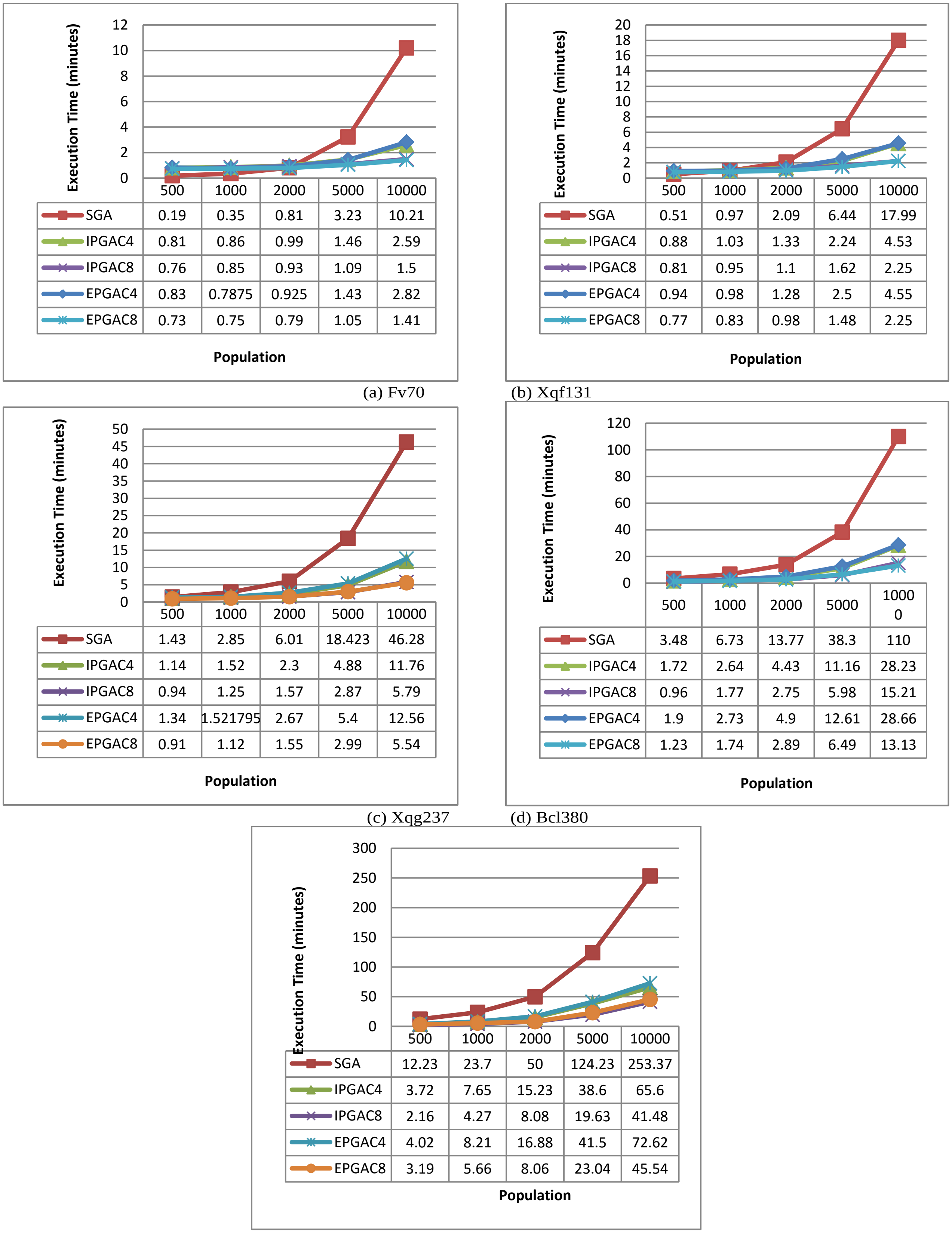
$$\text{Error } (\%) = \frac{\text{Best Solution} - \text{OptimumTSBLIB}}{\text{OptimumTSBLIB}} \times 100$$

TABLE I. MACHINES CONFIGURATION

Feature	Value
Architecture	64 bit
CPUs	4 cores
RAM	8 GB
Storage	500 GB
Operating System	Linux

TABLE II. CLUSTER CONFIGURATION EXPLOITED BY PGAS.

Name	Master nodes	Slave nodes	Total nodes
C4	1	4	5
C8	1	8	9



	500	1000	2000	5000	10000
SGA	0.19	0.35	0.81	3.23	10.21
IPGAC4	0.81	0.86	0.99	1.46	2.59
IPGAC8	0.76	0.85	0.93	1.09	1.5
EPGAC4	0.83	0.7875	0.925	1.43	2.82
EPGAC8	0.73	0.75	0.79	1.05	1.41

(a) Fv70

	500	1000	2000	5000	10000
SGA	0.51	0.97	2.09	6.44	17.99
IPGAC4	0.88	1.03	1.33	2.24	4.53
IPGAC8	0.81	0.95	1.1	1.62	2.25
EPGAC4	0.94	0.98	1.28	2.5	4.55
EPGAC8	0.77	0.83	0.98	1.48	2.25

(b) Xqf131

	500	1000	2000	5000	10000
SGA	1.43	2.85	6.01	18.423	46.28
IPGAC4	1.14	1.52	2.3	4.88	11.76
IPGAC8	0.94	1.25	1.57	2.87	5.79
EPGAC4	1.34	1.521795	2.67	5.4	12.56
EPGAC8	0.91	1.12	1.55	2.99	5.54

(c) Xqg237

	500	1000	2000	5000	10000
SGA	3.48	6.73	13.77	38.3	110
IPGAC4	1.72	2.64	4.43	11.16	28.23
IPGAC8	0.96	1.77	2.75	5.98	15.21
EPGAC4	1.9	2.73	4.9	12.61	28.66
EPGAC8	1.23	1.74	2.89	6.49	13.13

(d) Bcl380

	500	1000	2000	5000	10000
SGA	12.23	23.7	50	124.23	253.37
IPGAC4	3.72	7.65	15.23	38.6	65.6
IPGAC8	2.16	4.27	8.08	19.63	41.48
EPGAC4	4.02	8.21	16.88	41.5	72.62
EPGAC8	3.19	5.66	8.06	23.04	45.54

(e) Rbu737

Fig. 4. Execution Times Achieved by SGA and PGAs on the five TSP Datasets.

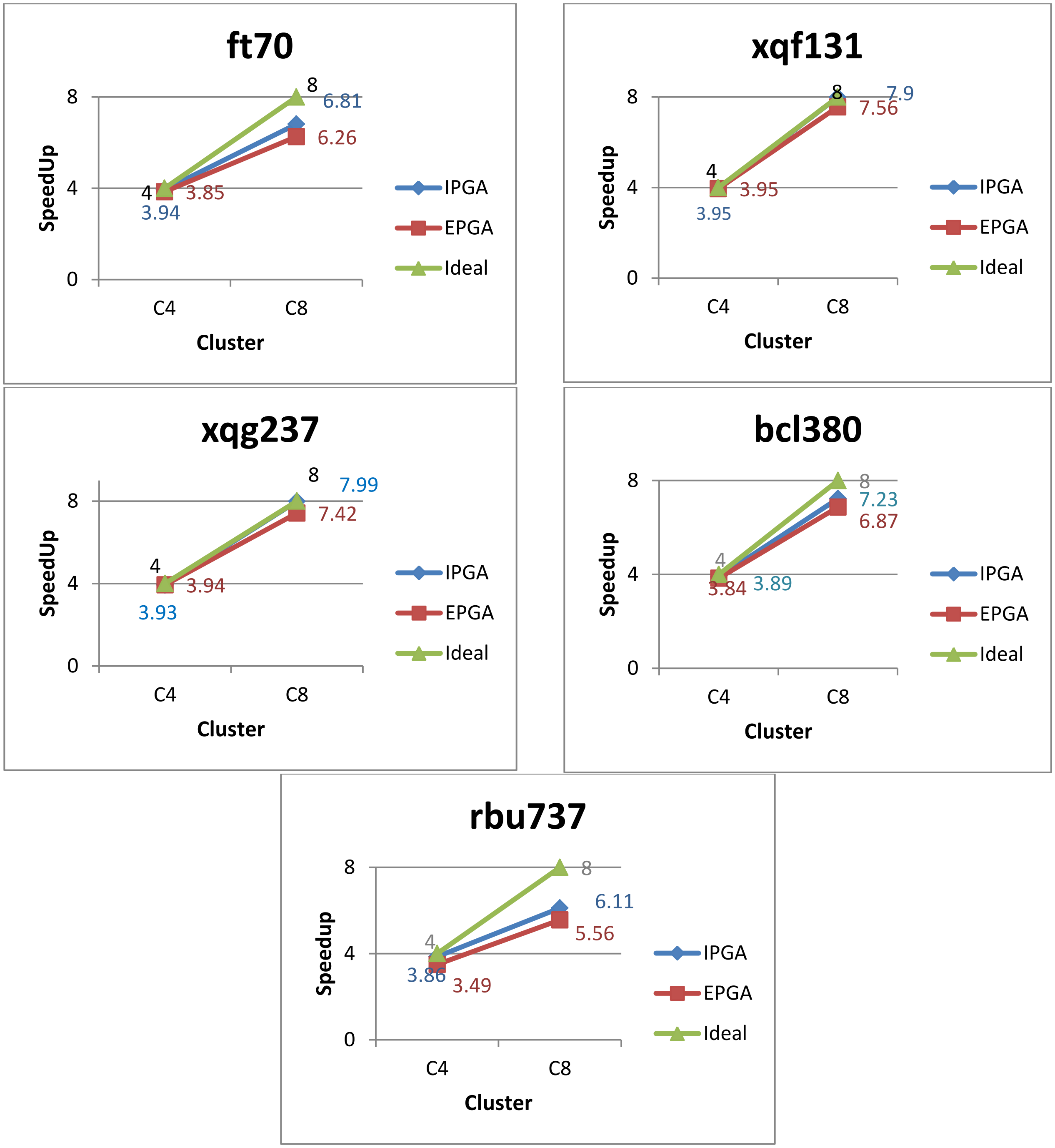
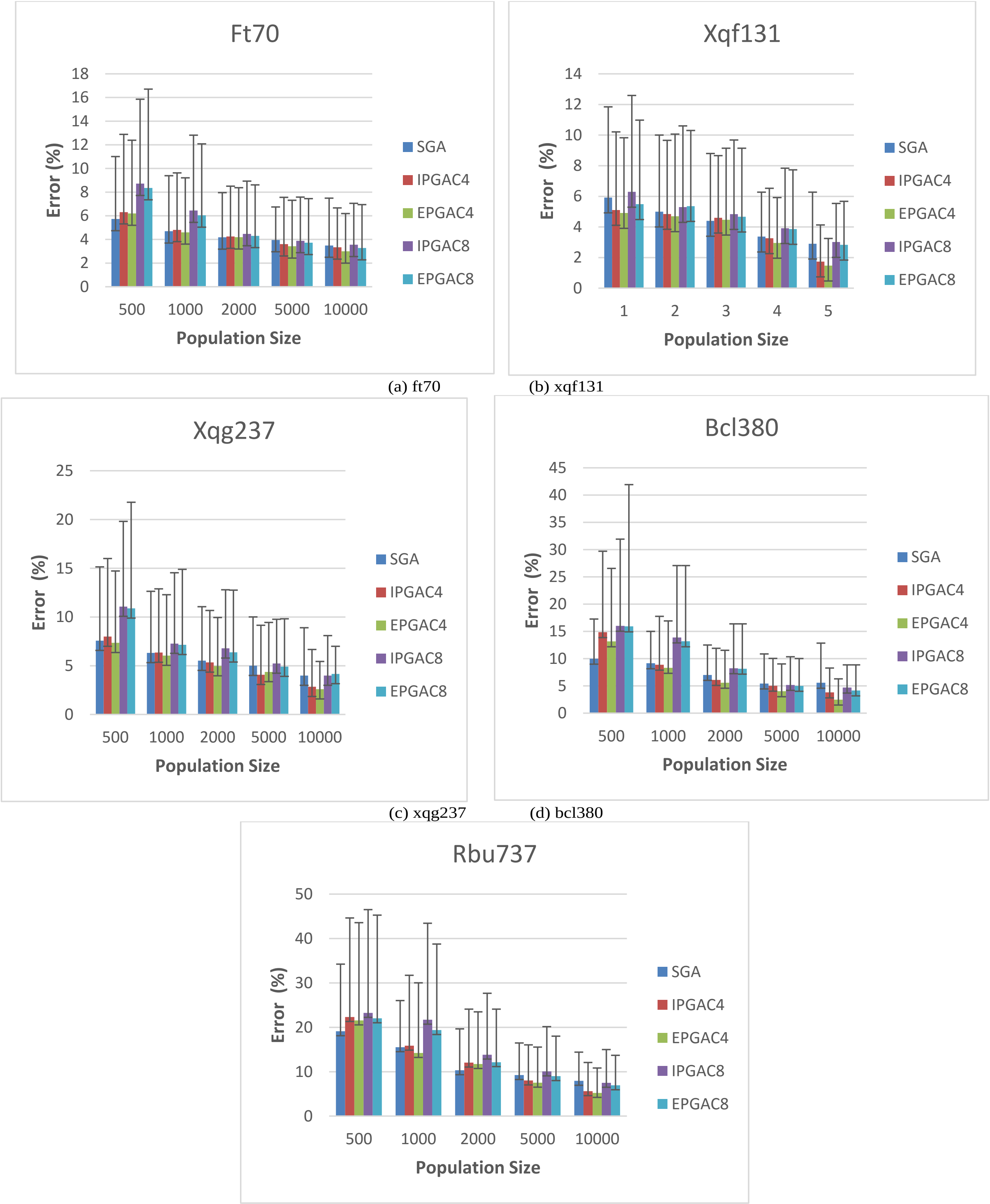


Fig. 5. IPGA and EPGA Speedup Per Dataset with Population Size 10000.

Fig. 6 reports the percentage of average solution accuracy achieved by SGA and PGAs on 20 runs on each algorithm and TSP dataset. We can observe from the figure that, the EPGA on 4-node cluster outperforms the SGA and IPGA for all considered datasets. Even if PGA obtain 'similar' solution accuracy as the SGA, the PGA outperforms the SGA in term of required time to get the same solution with different population size. We can observe also, as the population size increases, the PGAs performance improves rapidly in against to SGA. In Fig. 6(a), the EPGA in C4 cluster obtains better results and outperforms SGA and IPGA when the population size increases to 1k and above. The SGA obtains better results with small population sizes (i.e., 500 and 1000). This can be explained by the fact that the number of individuals on each island will be less than the original population (in case of 500, each island will have 125 and 63 individuals in C4 and C8 respectively), therefore the population diversity is also less than the original GA. This degrades search performance. In (b), the IPGA and EPGA in C4 cluster, outperform the SGA in all population sizes. The EPGA in C8 cluster outperforms the SGA after the population size increased to 2000 individuals, while the IPGA in C8 cluster remains the worst.



(a) ft70 (b) xqf131

(c) xqg237 (d) bcl380

(e) rbu737

Fig. 6. The Percentage of the Average Solution Quality Achieved by SGA and PGAs on the Five TSBLIB Instances.

In the xqg237 dataset, also, the EPGA obtains better results than all other models in all population sizes. In (d) and (e), we can observe that the SGA outperforms the PGA models in term of solution quality in the population sizes between 500 and 2000, but at the same time it takes very long execution time in against to PGAs. The PGAs obtained 'similar' and 'better' results faster than SGA but in larger population size. We can say that scaling up PGAs with large population tend to find better performance over the SGA within a reasonable time.

D. *Overhead*

The overhead time is the additional time other than the computation, due to communication and Hadoop platform tasks. The overhead is the reason that prevents the PGAs to have a speedup near the ideal on Hadoop platform. To measure the overhead for each PGA generation, we assign to each MapReduce job an initialization, computation, and finalization times for both Map and Reduce phases [16]. Fig. 7 shows the time measurement method for a MapReduce job.

- Map Initialization time is the time required to let the first Mapper start its computation.
- Map Finalization and Reducer Initialization time which is the time of the last mapper and the first reducer. Both of them are measured in the same way.
- Reducer Finalization time is the time of the last ending reducer.

In general, the overhead time in Hadoop corresponds to the sum of the overhead times of multiple jobs. Fig. 8 shows the mean computation and overhead times for each PGA on two datasets xqg237 and rbu737 in population size 2k. In Hadoop MapReduce, the overhead can be calculated using the sum of the overhead times of multiple jobs [16]. As we can see from the figure that, the overhead time for the IPGA and EPGA is the same and light in term of overhead time. That because we reduced the number of launched jobs during the PGAs execution time in order to control the overhead of HDFS accesses, which is limited to the migration phase only. From Fig. 8, we can observe that the overhead time is almost constant over the different jobs, and the map initialization phase takes longer time than the reducer initialization, that because in reducer phase nodes are already prepared to start reducer task when the Mappers are finishing. We also observed the overhead time independent of the dataset size.

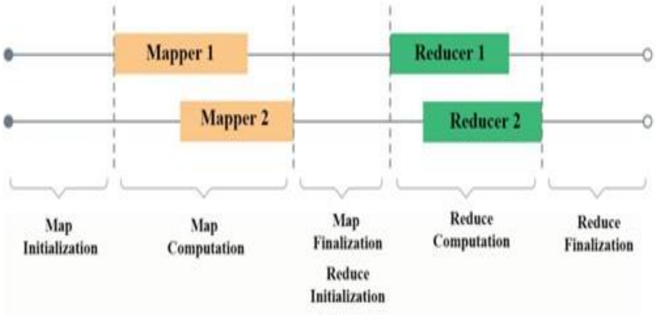
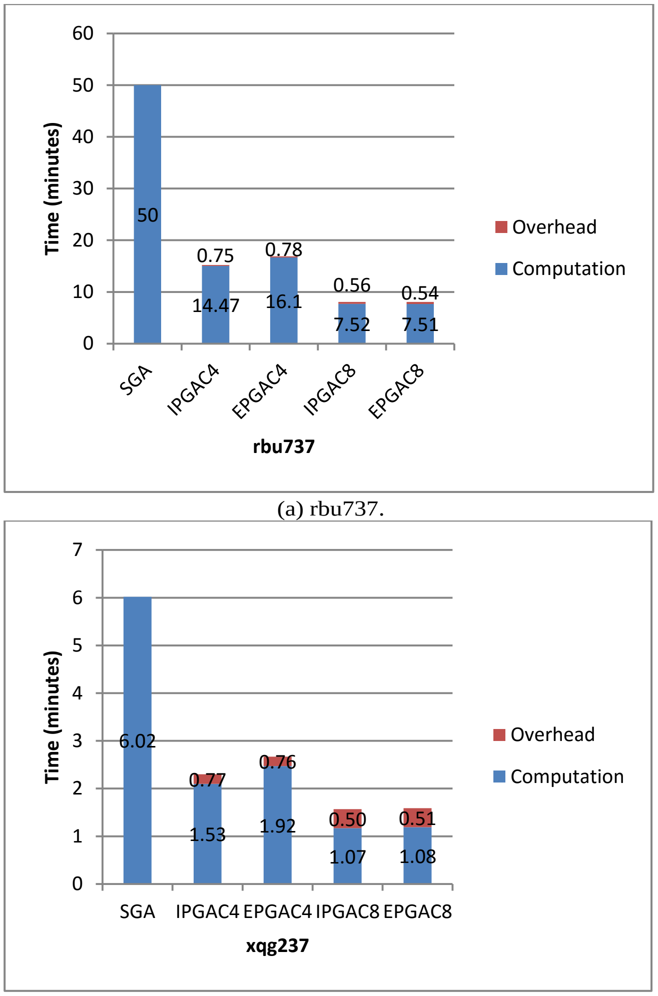


Fig. 7. The Time Measurement Method for Multiple Nodes.



(a) rbu737.

(b) xqg237.

Fig. 8. The Computation and Overhead Times for Each PGAs Model.

VI. Conclusion

In this paper, we designed a Hadoop MapReduce platform which is a general framework for implementing parallel genetic algorithms based on two models; the island model and the elite model. The traveling salesman problem is used as a benchmark in the experimental evaluation of those two Parallel Genetic Algorithms (PGAs).

We empirically assessed the effectiveness of those two PGA models in terms of execution time, speed up, overhead and solution quality by using five datasets form TSPLIB [24]. The datasets were chosen considering their different sizes in order to vary the execution times of the GAs. Additionally, varying population size and the cluster size were configured based on 4 and 8 parallel nodes.

We found that the PGAs find better solutions faster than Sequential Genetic Algorithm (SGA) when the problem size increases as well as when the population size increases. The EPGA outperforms the IPGA in term of the solution quality in a similar time for all the considered datasets and clusters.

We observed the effect of large population size, large populations (5k and 10k individuals) tend to find better solutions and need fewer generation periods to obtain good results which reduce the overall Hadoop overhead.

We also found that increasing the number of nodes in a cluster reduces the execution time. The use of the PGA models enabled to speed up the average execution time overall datasets with respect to SGA and tend to the ideal speedup value.

The overhead of HDFS access and communication is reduced in the PGAs since the number of operations performed on the datastore is limited to the migration phase only. However, Hadoop overhead may impair PGA solutions when executed on small problem instances.

We aim as a future work plan at comparing the performance of our model with an iterative MapReduce framework such as Haloop and Spark. Also, we aim to evaluate both parallel models by applying them to a challenging software engineering problem.

REFERENCES

[1] Y. J. Gong et al., “Distributed evolutionary algorithms and their models: A survey of the state-of-the-art,” Appl. Soft Comput. J., vol. 34, no. 2013, pp. 286–300, 2015.

[2] H. Ra and N. Erdoğan, “Parallel Genetic Algorithm to Solve Traveling Salesman Problem on MapReduce Framework using Hadoop Cluster,” Int. J. Soft Comput. Softw. Eng. [JSCSE], vol. 3, no. 3, pp. 380–386, 2013.

[3] N. E. A. Khalid, A. F. A. Fadzil, and M. Manaf, “Adapting MapReduce framework for genetic algorithm with large population,” in Proceedings - 2013 IEEE Conference on Systems, Process and Control, ICSPC 2013, 2013, no. December, pp. 36–41.

[4] J. Dean and S. Ghemawat, “MapReduce,” Commun. ACM, vol. 51, no. 1, p. 107, Jan. 2008.

[5] P. Sachar and V. Khullar, “Genetic Algorithm Using MapReduce-A Critical Review,” i-manager’s J. Cloud Comput., vol. 2, no. 4, pp. 35–42, 2015.

[6] E. Apostol, I. Băluţă, A. Gorgoi, and V. Cristea, “A Parallel Genetic Algorithm Framework for Cloud Computing Applications,” Pop F., Potop-Butucaru M. (eds). ARMS-CC 2014. Lect. Notes Comput. Sci. vol 8907. Springer, Cham, vol. 8907, pp. 113–127, 2014.

[7] W. Tom, Hadoop: The Definitive Guide. FOURTH EDITION The Definitive Guide STORAGE AND ANALYSIS AT INTERNET SCALE, 4th ed. O’Reilly Media, 2015.

[8] R. Li, H. Hu, H. Li, Y. Wu, and J. Yang, “MapReduce Parallel Programming Model: A State-of-the-Art Survey,” Springer, Int. J. Parallel Program., vol. 44, no. 4, pp. 832–866, 2016.

[9] F. Imeson and S. L. Smith, “A language for robot path planning in discrete environments: The TSP with Boolean satisfiability constraints,” in Proceedings - IEEE International Conference on Robotics and Automation, 2014, pp. 5772–5777.

[10] D. W. Huang and J. Lin, “Scaling populations of a genetic algorithm for job shop scheduling problems using mapreduce,” in Proceedings - 2nd IEEE International Conference on Cloud Computing Technology and Science, CloudCom 2010, 2010, pp. 780–785.

[11] K. Miclaus, R. Pratt, and M. Galati, “The Traveling Salesman Traverses the Genome: Using SAS® Optimization in JMP® Genomics to build Genetic Maps,” 2012.

[12] J. Singh and A. Solanki, “An Improved Genetic Algorithm on MapReduce Framework Using Hadoop Cluster for DNA Sequencing,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 5, no. 6, pp. 1238–1244, 2015.

[13] N. Bansal, A. Blum, S. Chawla, and A. Meyerson, “Approximation algorithms for deadline-TSP and vehicle routing with time-windows,” in Proceedings of the thirtysixth annual ACM symposium on Theory of computing, 2004, pp. 166–174.

[14] H. Bennaceur and E. Alanzi, “Genetic Algorithm For The Travelling Salesman Problem using Enhanced Sequential Constructive Crossover Operator,” Int. J. Comput. Sci. Secur., vol. 11, no. 3, p. 42, 2017.

[15] Z. H. Ahmed, “Genetic algorithm for the traveling salesman problem using sequential constructive crossover operator,” Int. J. Biometrics Bioinforma., vol. 3, no. 6, pp. 96–105, 2010.

[16] F. Ferrucci, P. Salza, and F. Sarro, “Using Hadoop MapReduce for Parallel Genetic Algorithms: A Comparison of the Global, Grid and Island Models,” no. x, pp. 1–33, 2017.

[17] C. Jin, C. Vecchiola, and R. Buyya, “MRPGA: an extension of MapReduce for parallelizing Genetic Algorithms,” in Proceedings - 4th IEEE International Conference on eScience, eScience 2008, 2008, pp. 214–221.

[18] A. Verma, X. Llorà, D. E. Goldberg, and R. H. Campbell, “Scaling Genetic Algorithms using MapReduce,” in In International Conference on Intelligent Systems Design and Applications (ISDA), 2009, pp. 13–18.

[19] A. Subasi and D. Keco, “Parallelization of genetic algorithms using Hadoop Map / Reduce,” SouthEast Eur. J. Soft Comput., no. June, pp. 127–134, 2012.

[20] T. Enomoto and M. Kimura, “Improving Population Diversity in Parallelization of a Real-Coded Genetic Algorithm Using MapReduce,” in Scientific Cooperations International Workshops on Electrical and Computer Engineering Subfields, 2014, no. August, pp. 234–239.

[21] R. Kondekar, A. Gupta, G. Saluja, R. Maru, A. Rokde, and P. Deshpande, “A MapReduce based hybrid genetic algorithm using island approach for solving time dependent vehicle routing problem,” in 2012 International Conference on Computer and Information Science (ICCIS), 2012, vol. 1, no. 2003, pp. 263–269.

[22] A. Rao, K. Hegde, K. Rao, IAnitha and Hegde, A. Rao, and S. K. Hegde, “Literature Survey On Travelling Salesman Problem Using Genetic Algorithms,” Int. J. Adv. Res. Eduation Technol., vol. 2, no. 1, p. 4, 2015.

[23] L. N. G. Sanchez, J. J. T. Armenta, and V. H. D. Ramırez, “Parallel Genetic Algorithms on a GPU to Solve the Travelling Salesman,” Difu100ci@, vol. 8, no. 2, pp. 79–85, 2014.

[24] A. Rohe, “VLSI Data Sets.” [Online]. Available: http://www.math.uwaterloo.ca/tsp/vlsi/index.html. [Accessed: 17-Nov-2018].

[25] G. Reinelt, “TSPLIB.” [Online]. Available: https://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/. [Accessed: 17-Nov-2018].

Twitter Sentiment Analysis in Under-Resourced Languages using Byte-Level Recurrent Neural Model

Ridi Ferdiana[1], Wiliam Fajar[2], Desi Dwi Purwanti[3], Artmita Sekar Tri Ayu[4], Fahim Jatmiko[5]
Department of Electrical Engineering and Information Engineering
Universitas Gadjah Mada, Yogyakarta, Indonesia[1, 2, 3, 4]
Microsoft Innovation Center, Universitas Gadjah Mada, Yogyakarta, Indonesia[5]

***Abstract*—Sentiment analysis in non-English language can be more challenging than the English language because of the scarcity of publicly available resources to build the prediction model with high accuracy. To alleviate this under-resourced problem, this article introduces the leverage of byte-level recurrent neural model to generate text representation for twitter sentiment analysis in the Indonesian language. As the main part of the proposed model training is unsupervised and does not require much-labeled data, this approach can be scalable by using a huge amount of unlabeled data that is easily gathered on the Internet, without much dependencies on human-generated resources. This paper also introduces an Indonesian dataset for general sentiment analysis. It consists of 10,806 twitter data (tweets) selected from a total of 454,559 gathered tweets which taken directly from twitter using twitter API. The 10,806 tweets are then classified into 3 categories, positive, negative, and neutral. This Indonesian dataset could help the development of Indonesian sentiment analysis especially general sentiment analysis and encouraged others to start publishing similar dataset in the future.**

Keywords—Sentiment analysis; under-resourced problem; Indonesian dataset; twitter

I. Introduction

Sentiment analysis is a problem of systematically identifying and studying personal information. This is commonly translated into the task of classifying polarity detection (thus this term is used interchangeably): Given a piece of written text, the problem is to categorize text into positive or negative classes or can be expanded to the ordinal classification problem. It assigns text to a value (e.g., Numbers from -2 to +2) instead of only positive or negative. There are some who think that polarity detection is not only related to the term sentiment analysis, polarity detection is only one subtask of the sentiment analysis process [1], [2]. However, this article uses the term sentiment analysis and polarity detection interchangeably as a focus on this task in this work.

Plenty of methods have been introduced to deal with sentiment analysis problem in previous studies. In general, the method can be either supervised or unsupervised. A lexicon-based approach is often used in unsupervised cases, where a list of words with their sentiment score is required to assign overall sentiment of a document. On the other hand, supervised machine learning techniques can also be considered to build sentiment analysis system because there is no such exact mapping between patterns of character in the text and the polarity of the sentiments (positive or negative). To produce a model from a series of data and let the computer to learn the patterns. There are several machine learning methods for classifying polarity detection: neural networks [3], [4], decision trees [5], support vector machines (SVM) [6], and naive Bayesian [7]. Feature pre-processing and extraction are carried out before classification, which requires large computing power.

Both machine-learning and lexical-based methods need extensive resources that are manually prepared. Lexical-based methods need sentiment lexicons, while machine-learning-based needs a lot of labeled data. This may be scarcely available to many languages, especially non-English languages such as Indonesian. Human-generated resources are expensive, which require much time and manual labor. This problem motivates us to ease the problem by adding a resource that may help other researchers to conduct research in this area and proposing a sentiment analysis system that leverages unsupervised approach, which minimizes the need of human-generated resources.

In this paper, it is proposed an unsupervised method for addressing the under-resourced problem in sentiment analysis for the Indonesian language. This article presents a methodology to use a byte-level self-supervised neural network to generate sentence representation in sentiment analysis in Indonesian, under the hypothesis that leveraging this method with an existing popular technique such as TF-IDF method will make improvements in this sentiment analysis classification performance.

Our main contributions are as follows:

- The use of unsupervised approach to minimize the under-resourced problem in the Indonesian language, particularly the byte-level recurrent neural model to generate a representation of sentences.
- To gather twitter dataset that contains 10,806 labeled samples and 454,559 unlabeled samples, hoping this would be one resource of doing evaluation benchmark when building a sentiment analysis system in the Indonesian language.

II. Related Work

This section overviews existing research on sentiment analysis, focusing on sentiment analysis in general, with emphasis on the Indonesian language.

The work conducted in this paper is sponsored by Microsoft Rinna

A. Feature Extraction

Feature extraction techniques are used to compress the data in a more compact way than the raw data such that the redundancy is removed while retaining relevant information [8]. Data patterns can be more easily discovered so this will ease the classification task. A good feature is required to have discriminatory properties, i.e., maximizing inter-class variability while minimizing intra-class variability. Machine learning systems can increase with the appropriate representation of features.

One of favorite feature in sentiment analysis is the Lexicon feature [9]. It uses a list of negative and positive words that are used to express the positive and negative sentiment. In the English dataset, the researcher can use SentiWordnet [10], or other similar databases. Based on Lexicon features, to determine the sentiments given by text, it is not necessary to train machine learning-based classifiers. it calculates the positive and negative polarity of text based on the occurrences of the positive and negative word using Pointwise Mutual Information (PMI) [11]. In this task, it may check whether the total value of the threshold is certain to determine its polarity. The Lexicon feature can be useful if have a complete list of words or can make a dictionary. But to make it takes a long time for manual work and may not be available in non-English languages. In addition, the N-gram feature is also popularly used by many works [1], [6] (a set of words / n-pair words). it counts the occurrence of words (1-gram or unigram) or n-pairs of words in the dictionary that have been determined to form the feature n-gram sentence.

Feature extraction techniques produce hand engineering features, i.e. the process of generating hand-crafted features is explicitly driven by predetermined algorithms. Designing such an algorithm takes time and requires a human expert. Therefore, there are several attempts to delegate the task of design extraction of this feature automatically to a computer. For classification, computers can determine for themselves which should be the best feature, considering raw data. This approach is called learning representation. Because computer data processing is increasing and more data is available, this is becoming popular in modern machine learning systems.

For text called word2vec, Mikolov et al. [12] proposed a method of learning representation that transforms words into multi-dimensional vectors. This transformation is carried out by a neural network encoder that is trained to predict the following words in the text. First, the neural network is initialized randomly and converts the word into a random vector. But once trained, encoders change the word vectors in a structure so that the words with similar meaning have a close distance and a pair of words that have a certain relationship will likely have the same distance as the other pairs of words that have the same relationship. This word embedding feature can be used in several specific NLP tasks such as sentiment analysis, text summarization and generating sentences that are given images.

To predict the preceding and succeeding sentence given a sentence, Kiros et al. [13] extends the success of the word insertion method by building sentence encoding by training neural networks. Inspired by these works, Radford et al. [14] proposed the learning of byte-level text representation. They propose an encoder that can produce multi-dimensional feature vectors from a sentence. These neural network encoders are also repeatedly trained to predict text that is not labeled and widely available on the internet. Instead of using word sequences as inputs, encoders are fed character by character. To predict the semantic polarity of text, encoders are sorted by a particular classifier machine learning.

B. Recent Workshops on Sentiment Analysis

SemEval is one of the challenges that has been held every year since 2013 [15]. In 2017, there were 48 teams that were successfully drawn by SemEval to be involved in the task of tweet sentiment analysis. In this task, participants will determine the sentiment value given by the tweet data. There are several techniques used, namely Logistic Regression, Random Forest, Maximum Entropy, Conditional Random Field, and Naif Bayes classifiers. SVM is more popular, and the best performing teams use such deep neural network as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). The top 5 teams for the English dataset use lexical, semantic features, dense word embedding, and the use of ensemble features. Available metadata for each tweet, such as the number of followers, user id, location, time zone, name, and a number of friends was not used by participants because they cannot increase their model performances. Analyzing and using effective metadata is very possible for future work.

SemEval continued to be held in 2018 [16]. A similar task is found in Task 1: Affect in Tweets, valence ordinal classification subtask. In the subtask, a participant is required to classify tweets into one of seven classes (-3, 2, -1, 0, +1, +2, +3) that represents the correct sentiment value. The best performing teams still used deep learning techniques such as CNN, LSTM, Gated Recurrent Unit (GRU), Bi-LSTM, and word embedding feature extraction methods combined with manually engineered features, i.e., sentiment and emoticon lexicons [17]. In the Arabic assignments, many teams use pre-processing techniques before doing the classification, such as stemming, lemmatization (MADAMIRA tool). By evaluating the result, it can be seen that although deep learning is interesting, performance can be improved by working together with hand engineering methods, which include feature pre-processing and extraction methods.

The use of non-English languages is still limited to several languages such as Spanish and Arabic. It may also be interesting to see other small languages such as Indonesian, Javanese, or Malay can be objects for the coming SemEval.

C. Challenges on Non-English Sentiment Analysis

Sentiment analysis of non-English texts has limited resources, such as Indonesian. Many unlabeled data available on the internet and labeled data are rarely available, so building an effective supervised machine learning system in non-English data can be challenging, especially if deep learning is used. In developing the Lexicon feature for sentiment analysis, a dictionary is needed in the form of a collection of negatives and positive sentiment words, which are not publicly available in Indonesian to the best of our knowledge.

The approach taken to build a lexicon dictionary can be done using automatic understanding in the available English lexicon database. Franky et al. [18] built an Indonesian lexicon dictionary by translating available English lexicons (Opinion lexicons, Harvard General Inquirer, SentiWordnet, and Bing Liu Opinions) using Google, Moses translations (statistical machine translation systems), and Kamus.net online dictionaries. Other works generally also use the availability of the English Lexicon database and conduct a mapping between English words and appropriate words in the language in which they work for the manufacture of non-English lexicon dictionaries [19], [20].

There are several works using the remote-controlled corpus to reduce the effort to build labeled datasets. Assigning labels to datasets is carried out remote monitoring using weak labels. For example, [21] forms a large number of labeled twitter datasets by setting tweets as positive if the text contains positive emoticons (such as ":)") and assigning it as negative if it contains any negative emoticon (such as “:(”). Author in [21] proposed a multiple language CNN for sentiment analysis, which is CNN that is trained with various corpus with different language to increase the number of data samples further, so it is able to handle multiple languages in a text. However, it does not perform well compared with CNN trained with single language dataset. Author in [26] proposes distant learning as an additional training set for Convolutional Neural Network for sentiment analysis. It is shown that the usage of distant learning increases performance by 5 percent.

By manually giving labels, some researchers decided to create their own non-English datasets. Franky et al. [18] built a labeled dataset by manually annotating 446 sentences originating from user reviews in several domains on the KitaReview website. Others [22] use a semi-supervised approach to sentiment analysis to overcome the scarcity of labeled data. They first made a small dataset of around 400 words of data labeled and collected a lot of non-labeled data (3 million). Furthermore, slowly labeled data is labeled using a classifier that is trained in a small labeled dataset.

III. Methodology

The article uses Multiplicative Long Short-Term Memory (MLSTM) Cell to build a recurrent neural network model. MLSTM cells are modified version of Long Short-Term Memory (LSTM) cells that are hybridized with Multiplicative Recurrent Neural Network [23]. MLSTM cells are observed to converge faster than LSTM during training [23]. The model considers an input sentence as a sequence of characters. Each character is encoded by a byte of Unicode encoding UTF-8. During training, the hidden state of the model is updated for each byte and the model predicts a probability distribution over the next possible character. The hidden state of the model encodes all information the model has learned from the first sequence and it provides relevant information to predict the future bytes of the sequence.

It is explored the optimal hyperparameter of the neural network model. it chose the embedding size of 128 and the RNN size of 4096. All of the states are assigned to 0 at the beginning, Adam method is used to train the neural network model with a learning rate of 5 X 10^{-4} that was decreased to zero over the training iterations. The model is trained with a dataset that contains tweets sentences. Because the training objective is to predict the next sequence of input, it does not need labeled data samples so the dataset can be easily gathered from the Internet. In this experiment, it used 454,559 unlabeled data gathered from the twitter API.

Once the recurrent neural network model is trained, the neural network model is used as a feature extractor of text input. The model processes an input byte by byte, forming its hidden state that is regarded as a multidimensional dense vector representation of the input sentence. The vector is then used along with a classifier such as SVM and trained with labeled data to create a sentiment prediction model.

A. Dataset

Supervised learning required a huge amount of labeled data, which is hard to come by. This is especially true for Indonesian dataset. Although gaining a lot of popularity, most of the research conducted in Indonesian sentiment analysis sometimes use only 1,000 to 5,000 data in their experiment.

The article chooses twitter as our dataset because twitter is one of the most popular sources for sentiment analysis data. Tweets consist of a maximum of 280 characters. And due to the nature of Twitter itself, most tweets only consist of text, unlike other social media such as Facebook or Instagram. The data gathering methods from twitter is also relatively easy. With twitter API, researchers can access not only tweets but also location, languages, user information, etc., which makes it easier to gather any specific data needed. There is also hashtag in twitter that makes data gathering process for sentiment analysis on a certain topic easier. All of these things on top of the other make Twitter our choice for a sentiment analysis dataset source.

This dataset in this work is originated from Twitter and taken with Twitter API between September and December of 2018. Each of the tweets has a maximum character of 140 from the gathered data of 454,559 tweets, 10,806 tweets were selected to be labeled and used. The example of the dataset can be seen in Table I.

TABLE. I. SAMPLE OF THE LABELED INDONESIAN DATASET

Sentiment	Indonesian Tweet
-1	lagu bosan apa yang aku save ni huhuhuhuhuhuhuhuhuhuuuuuuuuuuuuuu
-1	kita lanjutkan saja diam ini hingga kau dan aku mengerti tidak semua kebersamaan harus melibatkan hati
1	doa rezeki tak putus inna haa zaa larizquna maa lahu min na fadesungguhnya ini ialah pemberian kami kepada kamu
1	makasih loh ntar kita bagi hasil aku 99 9 sisanya buat kamu
0	yg selama ini nunggun video dari aku nih retweet yg mau full
0	barusan liat tulisan di belakang truk rela injek kopling demi kamu bisa shopping
-1	ku kira aku introvert ternyata karna nga punya uang aja jadi males ke mana
1	ya aku akan menjadi satu satunya bukan nomor satu tetapi satu satunya
-1	aku aja capek sama diriku sendiri apalagi kamu maaf ya '
0	aku nang kampus untuk ngopi ketemu wong2 podo kabeh takok e lapo mengubah sosyeti

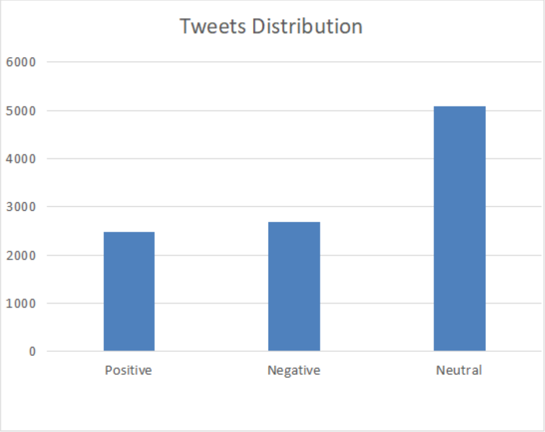


Fig. 1. The Distribution of the Sentiment in the Indonesian Dataset.

The selected tweets then labeled manually with three variables which is positive, labeled as 1, negatively labeled as -1, and neutral, labeled as 0. From the total of 10,806 tweets, 2482 are labeled as positive tweets, 2691 as negative tweets, and 5084 as neutral tweets; the distribution can be seen in Fig. 1.

There are 1:1:2 ratio of positive, negative, and neutral tweets, but considering the main purpose of this dataset is general sentiment analysis, it is concluded that the balance between each category is sufficient to be used even besides general sentiment analysis. The tweets are saved in CSV format with two columns. These tweets also have been lightly processed to remove noise so it can be conveniently used, the noise removed are symbols, URL links, username, and hashtag. The dataset can be downloaded as a common creative copyleft license at http://ugm.id/idsadataset

IV. Result and Discussion

It is conducted experiments to evaluate the effectiveness of the model for generating text representation from the byte-level recurrent neural network. It is shown comparison results between the proposed model and other typical sentiment analysis models: SVM classifier with TF-IDF features, and sentiment lexicons using AFINN [24]. The performance is evaluated using standard evaluation metrics: accuracy. Accuracy is defined as:

$$\text{Acc} = \frac{\text{true\_negatives} + \text{true\_positives}}{\text{true\_positives} + \text{false\_positives} + \text{false\_negatives} + \text{true\_negatives}}$$

To get sentiment value using AFINN, it is translated the sentiment lexicons dictionary from English to Indonesian using Microsoft translation service. Sentiment analysis is performed by cross-checking the string tokens (words, emojis) with the translated AFINN list and getting their respective scores.

TABLE. II. Effectiveness Comparison among our Model and other Typical Approaches

Method	Accuracy
1-gram TF-IDF vector	0.528
byte-level recurrent neural model	0.543
AFINN sentiment lexicon	0.455

Table II shows the results of our model on our labeled tweet datasets and TF-IDF features with an SVM classifier. TF-IDF representation provides 52.8 % of accuracy, while byte-level generated features give 54.3% of accuracy. There is 2.84 % improvement when using byte-level generated features compared to typical TF-IDF features. The result can be improved by concatenating the feature vectors of TF-IDF and the character level word embedding and making use of principal component analysis dimensionality reduction technique. AFINN sentiment lexicon methods give 45.48 % of accuracy.

V. Conclusion

In this work, it is proposed the use of byte-level recurrent neural networks with multiplicative long short-term memory cells for generating a representation of sentences, which are combined with a classifier (such as SVM) to generate a prediction of sentiment. The hybrid representation addition with sentiment lexicon could improve accuracy.

It cannot be said that the proposed methodology performance beat the state-of-the-art. On the other hand, state-of-the-art approaches, require a considerable amount of human work, which are labeled dataset and sentiment lexicon dictionaries. The proposed methodology is simple and does not rely on human-generated resources so it can be scalable to a larger dataset. However, it requires huge computational resources to conduct this methodology, as it has to process a huge amount of unlabeled data.

In the future, the research will aim to conduct experiments towards the following directions, in order to improve its performance: (a) improvement of the use pre-processing methods of text, (b) Apply the methodology into a larger dataset (e.g. contains millions of data samples).

Acknowledgment

The authors would like to thank you to Microsoft Rinna for their kind supports in providing grants and facilities for conducting this research.

References

[1] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, “New Avenues in Opinion Mining and Sentiment Analysis,” IEEE Intell. Syst., vol. 28, no. 2, pp. 15–21, 2013.

[2] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera, “Distinguishing between facts and opinions for sentiment analysis: Survey and challenges,” Inf. Fusion, 2018.

[3] X. Zhang and X. Zheng, “Comparison of Text Sentiment Analysis Based on Machine Learning,” 2016 15th Int. Symp. Parallel Distrib. Comput., pp. 230–233, 2016.

[4] J. Wehrmann, W. Becker, H. E. L. Cagnini, and R. C. Barros, “A character-based convolutional neural network for language-agnostic Twitter sentiment analysis,” 2017 Int. Jt. Conf. Neural Networks, pp. 2384–2391, 2017.

[5] Z. Rezaei and M. Jalali, “Sentiment analysis on Twitter using McDiarmid tree algorithm,” 2017 7th Int. Conf. Comput. Knowl. Eng. ICCKE 2017, vol. 2017-Janua, no. Iccke, pp. 33–36, 2017.

[6] Ike P. Windasari, F. N. Uzzi, and iman satoto Kodrat, “Sentiment Analysis on Twitter Posts: An analysis of Positive or Negative Opinion on Gojek,” in IEEE International Conference on Information technology, Computer, and Eletrical Engineering, 2017.

[7] T. Ghorpade and L. Ragha, "Featured based sentiment classification for hotel reviews using NLP and Bayesian classification," Proc. - 2012 Int. Conf. Commun. Inf. Comput. Technol. ICCICT 2012, pp. 1–5, 2012.

[8] S. Pasarate and R. Shedge, "Comparative study of feature extraction techniques used in sentiment analysis," 2016 Int. Conf. Innov. Challenges Cyber Secur., no. Iciccs, pp. 182–186, 2016.

[9] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," Comput. Linguist., 2011.

[10] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining SentiWordNet," Analysis, vol. 10, pp. 1–12, 2010.

[11] K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," in Proceedings of the 27th annual meeting on Association for Computational Linguistics -, 1989, vol. 16, no. 1, pp. 76–83.

[12] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Proc. Int. Conf. Learn. Represent. (ICLR 2013), pp. 1–12, 2013.

[13] R. Kiros et al., "Skip-Thought Vectors," 2015.

[14] A. Radford, R. Jozefowicz, and I. Sutskever, "Learning to Generate Reviews and Discovering Sentiment," Apr. 2017.

[15] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 Task 4: Sentiment Analysis in Twitter," Proc. 11th Int. Work. Semant. Eval., pp. 502–518, 2017.

[16] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," Proc. 12th Int. Work. Semant. Eval., pp. 1–17, 2018.

[17] K. Baktha and B. K. Tripathy, "Investigation of recurrent neural networks in the field of sentiment analysis," Proc. 2017 IEEE Int. Conf. Commun. Signal Process. ICCSP 2017, vol. 2018-Janua, pp. 2047–2050, 2018.

[18] Franky, O. Bojar, and K. Veselovská, "Resources for Indonesian Sentiment Analysis," Prague Bull. Math. Linguist., vol. 103, no. 1, pp. 21–41, 2015.

[19] A. Bakliwal, P. Arora, and V. Varma, "Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification," eighth Int. Conf. Lang. Resour. Eval., 2012.

[20] V. Perez-Rosas, C. Banea, and R. Mihalcea, "Learning Sentiment Lexicons in Spanish," Proc. Eighth Int. Conf. Lang. Resour. Eval., 2012.

[21] J. Deriu et al., "Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification," 2017.

[22] N. F. F. Da Silva, L. F. S. Coletta, and E. R. Hruschka, "A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning," ACM Comput. Surv., vol. 49, no. 1, pp. 1–26, 2016.

[23] B. Krause, L. Lu, I. Murray, and S. Renals, "Multiplicative LSTM for sequence modelling," 2016.

[24] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," CEUR Workshop Proc., vol. 718, pp. 93–98, 2011.

Predicting Return Donor and Analyzing Blood Donation Time Series using Data Mining Techniques

Anfal Saad Alkahtani[1], Musfira Jilani[2]
Collage of Information and Computer Sciences
Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia[1]
School of Computing, Dublin City University, Dublin, Ireland[2]

***Abstract*—Since blood centers in most countries typically rely on volunteer donors to meet the hospitals' needs, donor retention is critical for blood banks. Identifying regular donors is critical for the advance planning of blood banks to guarantee a stable blood supply. In this research, donors' data was collected from a Saudi blood bank from 2017 to 2018. Machine learning algorithms such as logistic regression (LG), random forest (RF) and support vector classifier (SVC) were applied to develop and evaluate models for classifying blood donors as return and non-return donors. The natural imbalance of the donors' distribution required extra attention and considerations to produce classifiers with good performance. Thus, over-SMOTE sampling was tested. Experiments of different classifiers showed very similar performance results. In addition to the donors return classification, a time series analysis on the donors dataset was also considered to find any seasonal variations that could be captured and delivered to blood banks for better planning and decision making. After aggregating the donation count by month, results showed that the number of donations each year was stable except for two discovered drops in June and September, which for the two observed years coincided with two religious periods: Fasting and Performing Hajj.**

Keywords—Classification; machine learning; time series analysis; blood donation

I. Introduction

Blood donation is an integral and essential part of the healthcare system. Without blood banks, many of the medical procedures that we otherwise take for granted could not take place. The modern lifestyle, ever-increasing mobility and accompanying higher accident rates, and incidences of natural and human-made disasters (such as wars, earthquakes, etc.) have led to an ever rising demand for blood transfusions [1]. A constant supply of blood is needed to help ensure that hospitals have access to enough blood to meet their current and future needs. One of the most important factors for a stable supply is the retention of donors, as return donors allow blood banks to not spend additional efforts and resources looking for new ones [2].

Kingdom of Saudi Arabia (KSA) is a large country with a total population of 33413660 inhabitants [3]. According to the Ministry of Transportation, 13221 car accidents were recorded in 2018 [4]. As a result, more blood banks in all regions were highly needed. Despite several campaigns aimed at promoting voluntary blood donations [5], statistics show that KSA is lacking the adequate number of volunteer blood donors [6].

In order to assist in overcoming these challenges, research on donors' data is critical. Machine learning and data mining methods can be applied to analyze the behavior of donors, allowing blood banks to build binary classification models for the return prediction of donors. However, in a real dataset of donors, there are few return donors compared with a high number of non-return donors. This imbalance in the donors' dataset introduces another problem to the binary classification, a condition where the model will consider too highly the majority class of non-return donor and ignores the minority class of return donor (even though the return donors are the primary class of interest). Thus, special considerations are required. Moreover, exploring seasonal variations or trends can help blood banks in planning and decision making. Time series analysis and forecasting methods are helpful to discover such knowledge.

This study was conducted on a real data set collected by the author from a public hospital in Saudi Arabia. Moreover, this is the first study applied in Saudi Arabia in the domain of data mining applications on blood donors.

In addition to the current section, this paper is divided into six sections. Section 2 starts with an overview of the history of blood donation and its system, then presents the previous works in return donors predicting and blood donation time series analysis. The data collection is shown in Section 3. Next, Section 4 describes all methodologies used in the study. Then, Section 5 presents the discussion. Finally, Section 6 summarizes the whole research in the conclusion.

II. Literature Review

A. Blood Donation System

The blood supply chain in general consists of three main roles: blood donors, transfusion agencies (such as hospitals), and blood centers (which attempt to coordinate the balance between supply and demand). The system can be described in three stages as shown in Fig. 1.

Stage 1: Blood Collection

Blood centers invite donors to donate blood in a number of different ways, such as recruitment campaigns, direct mailings and phone calls. The donor can be a volunteer, a replacement (a patient's relative or friend), or a paid donor. The blood collection could be for whole-blood, plasma, or platelet. The collection process is performed in government institutions, companies, and hospitals [7].

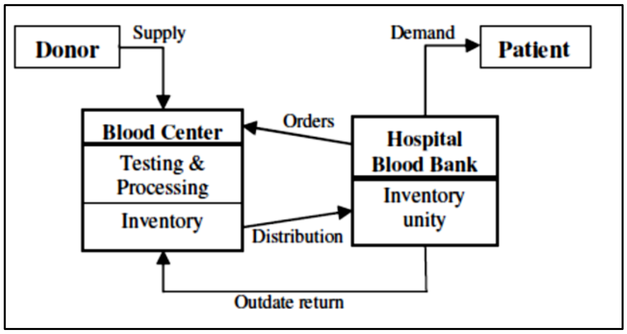


Fig. 1. Blood Supply Chain [7].

Stage 2: Blood test and processing

After collection, the blood centers apply serological and immunohematological tests on the collected blood bags. One of these tests is to determine the blood group of each bag, which can be one of the four different blood group (A, B, AB, and O) and is either Rh-positive or Rh-negative. Then, the bags are sent to processing units to extract and store blood components [7].

Stage 3: Blood distribution

Requests from hospitals to fill their blood bank needs are met by centers releasing and distributing blood components [7].

B. Blood Donation around the World and in Saudi Arabia

Blood centers all over the world are suffering a high shortage in the blood supply. Moreover, the demand for blood replacement is continually increasing in all countries [8]. The American Red Cross states that there is someone in need of blood every two seconds. However, the blood centers rely on volunteer donors, and most do not return to donate again. Unfortunately, not all donors are eligible, since a lot of donated blood bags are rejected after the test stage [9]. Another factor causing a decline of donation is that many returning donors are aging, whereas younger donors are rare [8].

Although the United States has a strong coordinated effort between the government and the Red Cross, the blood supply remains below the demand. In the U.S., the daily need of blood donors is around 44000 [10]. According to the Blood Transfusion Service in Northern Ireland, the yearly need of new donors is 10000 donors [10]. A study by Gibbs and Corcoran reported that “80% of developing countries depend totally or partially on replacement donors, 15% on voluntary/non-remunerated and 25% on paid donations” [9].

The KSA is a large country with a total population of 33413660 inhabitants [3]. According to the Ministry of Transportation, 13221 car accidents were recorded in 2018 [4]. As a result, many blood banks in all the regions are highly needed. At the Fourth World Medical Conference on Blood Transfusion organized at Prince Sultan bin Abdul-Aziz Center for Science and Technology (SITEC), the assistant undersecretary of the Ministry of Health for Laboratories and Blood Banks stated that there are a series of 6 central blood banks and 18 major blood banks provided for health services around the Kingdom. He stated that 60% of the need in the blood banks in the Kingdom is usually covered by the donation of relatives and friends [6]. The voluntary donation of blood meets only 40% of the total need, which is too low, as he would like to see that closer to 100% [6].

The cycle of blood donation in the KSA is started by the donor and ends with the transfusion to the patient. It is essentially a hospital-based procedure, managed by the blood banks of the individual hospitals. Thirty years ago, importing a blood supply was stopped, and the Kingdom decided to depend entirely on local blood donations [11]. Currently, the source of blood donations is a combination of mostly replacement donors and a growing number of voluntary donors. The latter source is expanding rapidly through better operations by blood bank managers [11]. Previously, participation in blood donation was less than satisfactory among the Saudi public. This was probably due to a poor awareness of the importance of blood donation. However, this is changing as a result of better communications between the hospital blood banks and the general public [12] [13].

C. Toward Blood Bank Analysis using Data Mining Techniques

As a result of the increased global demand for blood, there is a serious need to keep an adequate supply of blood that is readily available [7]. Finding a way to recruit new donors and encourage previous donors to return is a major challenge for blood centers. In order to address this challenge, many studies have been conducted to apply different data mining tasks in the blood donation field. For instance, predicting return donors and forecasting the number of donors over a short time interval. This section reviews some of the most relevant research regarding these tasks.

1) Return donors prediction: Most of the previous studies applied logistic regression to donor demographic information in order to predict returning donors and to understand the underlying factors affecting this prediction [14][15][16].

A study on the REDS donors' dataset, which had been collected by six blood bank centers around the U.S. from 2003 to 2004, was conducted along with a survey of the donors. The donors' dataset contained information of donation dates, donation status (first-time versus repeat), donation frequency, donation type, and other demographic characteristics (age, sex, and race). The class attribute (1: return donor, 0: non-return donor) was defined as 1 if the donor returned to donate during one year after completing the survey, and 0 otherwise. A binary logistic regression was used to determine the most significant predictive attributes ($p< 0.05$). The study found that predicting donor return was highly dependent on the donation frequency, the convenience of the donation place, and a good donation experience [14].

A study on demographic data, frequency of return, and the time interval between donations was conducted by collecting this data from the Shahrekord Blood Transfusion Center for five years. To conduct this study, researchers created a list of first-time donors for a single year (2008-2009), then the additional variables of frequency of return, the time interval between donations, and the class attribute return to donation (1: return at least once, otherwise 0: non-return) were collected for the next four years. The three response variables (return to donation, frequency of return and the time interval between donations) were analyzed using logistic regression, negative binomial regression, and Cox's shared frailty model, respectively. The results of the logistic regression showed that donor return was mainly affected by sex, weight, and career [15].

In the U.K., the National Health Service Blood and Transplant (NHSBT) agency conducted a study on a dataset of volunteer donors of whole blood for two years (2010-2011) that included these demographic data: donor ID, postcode, age, ethnicity, donation date, donation type, donor status flag (new or repeat), and most recent previous donation date. The class attribute for return prediction was defined as return donor if the interval between the last donation date and the most recent previous donation date was no more than two years, non-return donor otherwise. Notice that to make sure that first-time donors had enough time to donate again, the research removed all first-time donors that donated in the last six months of the study period. To analyze these features, a multivariate logistic regression was used and showed that donor return varied mainly by geographic location, age, and gender [16].

Other studies relied on survey data rather than blood donation datasets to analyze the factors affecting donors return, with more attention on geographical, educational, and awareness factors [17][18]. Generally, they found that the probability of repeating donation was affected by gender, age, education level, awareness of the donation time, and donor living area.

In addition to logistic regression, other machine learning techniques have been used to predict donor return. For instance, a Classification and Regression Tree (CART) classifier applied to public blood donation data available on the UCI Machine Learning Repository provided a high classification accuracy with 99% precision [19]. In another study with the same dataset, the ANN and SVM algorithms were used for classification with results of ANN sensitivity (65.8%), ANN specificity (78.2%) SVM sensitivity (68.4%), and SVM specificity (70.0%) [20].

2) Analyzing and forecasting blood donation: Previous studies in this area of blood donation are few. One similar study on red blood cell (RBC) transfusion by the blood bank of the Hospital Clinic of Barcelona investigated three univariate time-series methods for forecasting the monthly demand by analyzing the RBC demand dataset during a 15-year period (1988-2002). The performance of an autoregressive integrated moving average (ARIMA), a Holt-Winters exponential smoothing model, and a neural-network-based model were compared. The results indicated the excellence the ARIMA and exponential smoothing models for short-time forecasting [21].

The Hacettepe University Hospitals Blood Center conducted a study of investigating donor arrival patterns in order to determine the required workforce for each day of the week and also within a single day. The dataset contained 1095 records, each representing a single day during three years (2005-2007). Each day (record) was described by 20 attributes: year (1-3), month (1-12), day-of-month (1-31), day-of-week (1-7), and 16 attributes for a one-hour time period of the day (08:00- 24:00) and containing the arrival rate during that hour. The study applied clustering followed by classification methods (rather than a time-series analysis), and considered two donor arrival patterns: daily arrival patterns within the week, and hourly arrival patterns within the day. The results found that the arrival rates to the blood center varied based on ten distinct hourly patterns found within three identified daily patterns (Monday-Thursday, Friday, and Saturday-Sunday) [22].

III. Data Collection

Data of blood donors had been officially collected from a public Saudi hospital for a period of two years from 2017 to 2018 across various ages, blood types, genders and nationalities. It contains the date of the first donation for each donor and all dates of his/her subsequent donations during the period under study. Table I presents the dataset attribute descriptions.

A. Data Preprocessing

The major steps involved in data preprocessing are feature generation and data cleaning. Feature generation creates new attributes from the original data that can help the modeling. Data cleaning involves detecting outliers and handling noise and missing values.

1) Feature engineering

Three features were added as follows:

- *Last donation date*: This was extracted from the donation date feature by saving all donation dates for a donor during the two years on a sorted list (*DonationDatesList*), then retrieving the last element in this list.
- *Period in months:* This was calculated from the first donation date and last donation date.
- *Class label:* A donor was considered as return donor (1) if he/she donated two or more times within the period under study. Otherwise, the donor was non-return (0). This was calculated by using the length of *DonationDatesList* for each donor. The class is set to 1 if the length is 2 or more, 0 otherwise. Table II shows the total number of instances in each class after adding the attribute to the dataset.

TABLE. I. Dataset Attributes of Blood Donor Data

Attribute name	**Attribute type**	**Explanation**
ID	Integer	Donor id
Blood Type	Categorical	The blood type: A+, A-, B+, B- ,AB+,AB-,O+,O-
Gender	Categorical	Male (M), Female (F)
Age	Continuous	Range from 18 to 66
Donation count	Discrete	Total donations, from first to last donation dates
First donation date	Date	Date of first donation
donation date	Date	Date of each donation
Nationality	Categorical	Donor's nationality
City	Categorical	Donor's city

TABLE. II. NUMBER OF INSTANCES IN EACH CLASS

Class	Total
0	33018
1	1973
Total	34991

However, donors who donated once and for the first time in 2018 (14000 donors) had been removed from the analysis to ensure that all first-time donors had at least 1 year of follow-up time in which to make a subsequent donation. The reason behind choosing a 1-year follow up period is that the donors who had donated for the first time in 2017 returned in a period mean of 13 months. Among the remaining 21080 donors, 1945 donors were return-donors and 19135 were non-return donors.

2) Data cleaning

- Outliers' detection: There were some outliers in *Age*, *Period in months*, and *Donation count* attributes, as will be discussed later in the analysis section. However, such outliers are valuable to the analysis and will not be removed or changed.
- Missing values: There were missing values in two attributes, which were handled as follows:
  - The nationality variable had one missing value, which was replaced by KSA, as it was the majority value of the variable (99.99%).
  - The city variable had 44% with missing values. However, since Riyadh was the value for 99.99% of the rest, this variable was deleted.

After the preprocessing, the dataset was ready for analysis with the attributes listed in Table III.

TABLE. III. FINAL BLOOD DONOR DATASET

	Attribute name	**Attribute type**	**Explanation**
1	ID	Integer	Donor id
2	Blood Type	Categorical	The blood type: A+, A-,B, B-,AB+,AB-,O+,O-
3	Gender	Categorical	Male (M), Female (F)
4	Age	Continuous	Range from 18 to 66
5	Donation count	Discrete	Number of donations from first to last donation date
6	First donation date	Date	Date of first donation
7	Last donation date.	Date	Date of last donation
8	Nationality	Categorical	Donor nationality
9	Period in months	Discrete	The period in months from first to last donation date
10	Class	Binary	1 if return donor , 0 if non-return donor

IV. METHODOLOGY

This section illustrates the implementation of both classification and time series analysis. For classification, the impact of the imbalance problem in the return donors' prediction is presented by testing the selected classifiers before and after imbalance handling. Classifiers are evaluated in term of performance. Then, the blood donation time series analysis and forecasting are also discussed.

A. Predicting Return Donor

For the classification task that aimed to predict the donors' return, the implementation was conducted in two parts. First, imbalance in the donors' dataset was ignored and three classifiers were built: logistic regression (LG), random forest (RF), and support vector classifier (SVC). Second, the dataset was process while handling the imbalance by over-SMOTE sampling.

In all the constructed models, the following three steps were applied:

- Data Splitting: The data was split into a training set and a testing set. The training set was used to build and validate the models, while the testing set was treated as the new unseen data for evaluation.
- K-fold Cross Validation: A 5-fold cross-validation procedure with the training set was used to train and validate the models.
- Hyper-parameter optimization: A random search approach with cross-validation was used for tuning the hyper-parameters of the models to find the best combination.

The testing results of the three models before handling imbalance problem are shown in Table IV. The low recall of LG and SVC (around 30%) indicated that the models were suffering from an unrecognized positive class, i.e. many return-donors were predicted as non-return donors. On the other hand, in the recall of RF (65%) was higher which demonstrates that RF was better than LG (32%) and SVC (30%) in predicting the minority positive class (return-donor) correctly.

The LG, RF, and SVC models were constructed again, but after handling the imbalance with the Synthetic Minority Over-Sampling Technique (SMOTE). SMOTE was applied to the training set to create synthetic observations of returning donors. The testing results of the three models are shown in Table V.

TABLE. IV. TEST RESULTS FOR ALL CLASSIFIERS WITHOUT HANDLING THE IMBALANCE PROBLEM

Test	*LG*	*RF*	*SVC*
Accuracy	92.92	93.70	92.90
Precision	72.54	64.23	75.11
Recall	32.91	65.83	30.07
F1	45.28	65.02	42.94
AUC	65.85	81.13	64.54

TABLE. V. TEST RESULTS WITH RESAMPLING OVER-SMOTE

Test	LG	RF	SVC
Accuracy	92.67	92.64	92.61
Precision	55.14	55.04	54.95
Recall	94.48	94.12	93.77
F1	69.63	69.46	69.29
AUC	93.49	93.31	93.13

The three classifiers show a dramatic increase in recall after handling the imbalance by using over-sampling-SMOTE, from around 30% to 94 % in LG and SVC, and from 65% to 94% in RF. The improved recall demonstrates that the models are better at considering the minority positive class (return-donor) after introducing a balance in the dataset. Moreover, the differences between the LG, RF and SVC are very small for both F1 and AUC, where LG gives the best results.

B. Time Series Analysis and Forecasting

Time series analysis and forecasting from the donors dataset were applied to find any seasonal variations in blood donation that can be identified and communicated to blood banks for better planning and decision-making processes.

1) Blood donation time series analysis: The donation date attribute was converted to a date-time object. The dataset was then sorted by date and aggregated on a monthly basis with the *Resample* function. The monthly blood donation is shown in Fig. 2.

From the monthly blood donation plot over the 2-year period, it is possible to note that with the exception of June in each year, the mean shows some consistency. Moreover, the seasonality is pronounced as there are yearly drops in the month of Ramadan (roughly, June, for the period) and to a lesser extent on the Haj month (roughly, September, for the period). These two months match two religious periods (Fasting and Performing Hajj) and also two Islamic Eids, when people in the KSA have celebrate the holidays. During Ramadan, blood donation is not allowed during the fasting time unless as a necessity, e.g. replacement donors.

In order to see the series components, the series was decomposed into trend, seasonality, and residuals as shown in Fig. 3. The time series is clearly stationary, as the level of the series stays roughly constant over time, and the variance of the series appears roughly constant over time.

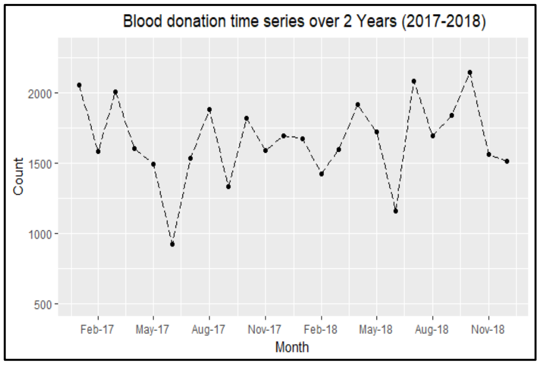


Fig. 2. Blood Donation over 2 Years.

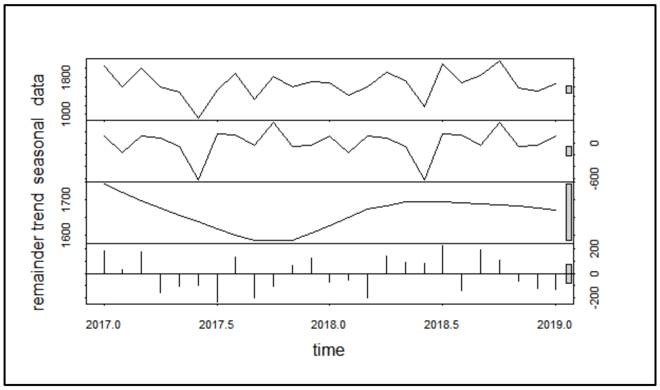


Fig. 3. Blood Donation Time Series Components.

2) Blood donation time series forecasting using ARIMA: Since the series appeared stationary, there was no need to apply any differences. The *d* value was set to 0. In order to determine the order of *p* and *q*, the ACF and PACF plots of the original time series were used to check the autocorrelation.

Since there was no significant lag in both ACF and PACF as shown in Fig. 4, the values of *p* and *q* were equal to 0. The final ARIMA model was ARIMA (0, 0, 0), which indicated that the series was a random walk and could not be predicted.

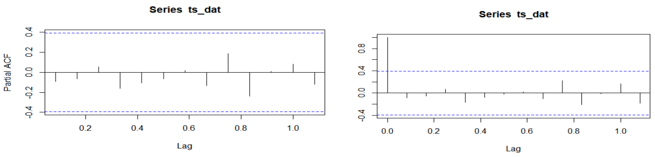


Fig. 4. Auto and Partial Correlation Functions of Blood Donation Time Series.

V. DISCUSSION

The dataset suffers from a limitation in attributes. More attributes can be collected to support the blood center in the prediction task to determine the return donor characteristics, for instance, profession, educational level, weight, neighborhood and reason for donation (volunteer or replacement donor).

As Recommended, hospitals can make use of blood donor's history to offer incentives for returning donors, such as flexible file opening and priority in appointments for them and their families. This would build strong and long-lasting relationships between return donors and the hospitals blood centers. Moreover, blood bank centers should target younger groups, especially in universities. For example, there are more than 40 universities in KSA that average more than 50.000 students each.

VI. CONCLUSION AND FUTURE WORK

In order to predict returning donors, data from donors to a Saudi blood center had been collected over two years (2017-2018). Machine learning algorithms were used to build binary classifiers that were able to predict return-donors with an imbalanced donors' distribution, where few instances belonged to the return-donor class and the majority instances were non-return donors. To illustrate the impact of imbalance problem in the prediction performance, all classifiers were tested before and after handling the imbalance. Experiments of different classifiers showed very similar performance results.

Moreover, the time series analysis of the monthly donation count explored stable donations over the year with two significant drops occurring during two religion periods, Fasting and Performing Hajj.

For future work, survival analysis and modeling can be applied to predict when the user will return for donation. Moreover, mixed models could be used to find the interaction between variables in the donors' dataset.

References

[1] T. Santhanam and S. Sundaram (2010). "Application of CART algorithm in blood donors classification," Journal of Computer Science, vol. 6 no. 5, p. 548.

[2] B.M. Masser, T.C. Bednall, K.M. White, and D. Terry (2012) "Predicting the retention of first-time donors using an extended theory of planned behavior," *Transfusion*, vol. 52, no. 6, pp. 1303-1310. Available at: 10.1111/j.1537-2995.2011.03479.x.

[3] "The total population in 2018," General Authority for Statistics. Available at: https://www.stats.gov.sa/en/indicators/1 [Accessed: February 13, 2019].

[4] "Ministry of Transport: The number of deaths of traffic accidents in the Kingdom's roads decreased by 33% in 2018," spa.gov.sa. Available at: https://www.spa.gov.sa/home.php?lang=en. [Accessed: February 13, 2019].

[5] M. Alam and D.B.D. Masalmeh (2004). "Knowledge, attitudes and practices regarding blood donation among the Saudi population," *Saudi Medical Journal*, vol. 25 no. 3, pp. 318-321.

[6] Obeed, M. (2018). "Al-Khashan: 'Health' is moving to encourage citizens to donate blood voluntarily," Sabq Newspaper Online Edition. Available at: https://sabq.org/4FL2xm [Accessed 3 Oct. 2018].

[7] O. Filho, W. Cezarino, and G. Salviano (2012). "A decision-making tool for demand forecasting of blood components," IFAC Proceedings, vol. 4 no. 6, pp. 1499-1504. Available at: 10.3182/20120523-3-ro-2023.00201.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[8] B. Sojoka and P. Sojka, "The blood donation experience: Self-reported motives and obstacles for donating blood," *Vox Sanguinis*, vol. 94, no.1, pp. 56-63. Available at: 10.1111/j.1423-0410.2007.00990.x.

[9] C. Holly, S. Balegh, and B. Ditto (2011). "Applied tension and blood donation symptoms: The importance of anxiety reduction." Health Psychology, vol. 30 no. 3, pp. 320-325. Available at: 10.1037/a0022998.

[10] M. Anwer, S. Ul Fawwad, S. Anwer and A. Ali (2018), "Attitude toward blood donation among medical and nonmedical students across Karachi,", *Asian Journal of Transfusion Science*, vol. 10, no. 2, p. 113. Available at: 10.4103/0973-6247.187937.

[11] A. Abdel Gader, F. Al Gahtani, A. Ramadan, A. Osman, M. Farghali and A. Al-Momen (2011) "Attitude to blood donation in Saudi Arabia," *Asian Journal of Transfusion Science*, vol. 5 no. 2, p. 121. Available at: 10.4103/0973-6247.83235.

[12] S. Hadi Alharbi et al. (2018). "Assessment of levels of awareness towards blood donation in Saudi Arabia," *AIMS Public Health*, vol. 5 no. 3, pp. 324-337. Available at: 10.3934/publichealth.2018.3.324.

[13] S. Elsafi, M. Al Zahrani, and E. Al Zahrani (2015). "Awareness and practice of blood donation by college students in Dhahran, Saudi Arabia," *ISBT Science Series*, vol. 10 no. 1, pp. 11-17. Available at: 10.1111/voxs.12172.

[14] K.S. Schlumpf, et al. (2008). "Factors influencing donor return," *Transfusion*, vol. 48 no. 2, pp. 264-272.

[15] S. Kheiri and Z. Alibeigi (2015). "An analysis of first-time blood donors return behaviour using regression models," *Transfusion Medicine*, vol. 25 no. 4, pp. 243-248.

[16] S. Lattimore, C. Wickenden, and S. Brailsford (2014). "Blood donors in England and North Wales: Demography and patterns of donation," *Transfusion*, vol. 55 no. 1, pp. 91-99. Available at: 10.1111/trf.12835.

[17] T. Volken, C. Weidmann, T. Bart, Y. Fischer, H. Klüter, and P. Rüesch (2013). "Individual characteristics associated with blood donation: A cross-national comparison of the German and Swiss population between 1994 and 2010," *Transfusion Medicine and Hemotherapy*, vol. 40 no. 2, pp. 133-138.

[18] W.I. Mauka, M.J. Mahande, S.E. Msuya, and R.N. Philemon (2015). Factors associated with repeat blood donation at the Northern Zone Blood Transfusion Centre in Tanzania," Journal of Blood Transfusion, vol. 2015.

[19] T. Santhanam and S. Sundaram (2010). "Application of CART Algorithm in Blood Donors Classification," *Journal of Computer Science*, vol. 6 no. 5, pp. 548-552. Available at: 10.3844/jcssp.2010.548.552.

[20] M. Darwiche, M. Feuilloy, G. Bousaleh, & D. Schang, D, "Prediction of blood transfusion donation," presented at Research Challenges in Information Science (RCIS), 2010 Fourth International Conference (2010, May).

[21] A. Pereira (2004). "Performance of time-series methods in forecasting the demand for red blood cell transfusion," *Transfusion*, vol. 44 no. 5, pp. 739-746. Available at: 10.1111/j.1537-2995.2004.03363.x.

[22] M. Testik, B. Ozkaya, S. Aksu and O. Ozcebe, "Discovering Blood Donor Arrival Patterns Using Data Mining: A Method to Investigate Service Quality at Blood Centers," *Journal of Medical Systems*, vol. 36 no. 2, pp. 579-594. Available at: 10.1007/s10916-010-9519-7.

Modified Farmland Fertility Optimization Algorithm for Optimal Design of a Grid-connected Hybrid Renewable Energy System with Fuel Cell Storage: Case Study of Ataka, Egypt

Ahmed A. Zaki Diab[1,*], Sultan I. EL-Ajmi[2], Hamdy M. Sultan[3], Yahia B. Hassan[4]
Electrical Engineering Department, Faculty of Engineering, Minia University, 61111 Minia, Egypt[1, 2, 3]
Department of Electrical and Electronic Engineering, Kyushu University, Fukuoka 819-0395, Japan[1]
Electric Engineering Department, Higher Institute of Engineering, 61111 Minia, Egypt[4]

***Abstract*—In this paper, a Modified Farmland Fertility Optimization algorithm (MFFA) has been presented for optimal sizing of a grid connected hybrid system including photovoltaic (PV), wind turbines and fuel cell (FC). The system is optimal designed for providing a clean, reliable and affordable energy by adopting hybrid power systems. This system is very important for countries looking to achieve their sustainable development goals. MFFA is proposed in order to reduce the processing implementation time. The optimization method depends on the high reliability of the hybrid power supply, small fluctuation in the injected energy to the grid and high utilization of the wind and solar complementary characteristics. Moreover, MFFA is applied to minimize the cost of energy while satisfying the operational constraints. A real case study of a hybrid power system for Ataka region in Egypt is introduced to evaluate the performance of the proposed optimization method. Moreover, a comprehensive comparison between the proposed MFFA optimization technique and the conventional Farmland Fertility Algorithm (FFA) has been presented to validate the proposed MFFA.**

***Keywords*—Photovoltaic; wind; fuel cell; renewable energy; hybrid energy system; modified farmland fertility optimization**

I. Introduction

Recently, fossil fuels comprise the major energy sources in many parts all over the world. In fact, these fossil fuels can be exhausted and usually negatively affect the environment while they are being utilized into useful forms of energy [1]. Taking into account the drawbacks besides using fossil fuels in energy application, there is an urgent need to find cost-effective, reliable and clean alternative sources of energy. Renewable sources of energy such as PV, wind turbine and hydropower are the most attractive candidates for clean energy generation especially in low scale and isolated areas [2]. Each of these renewable energy sources has its unique character and principle of operation, which make it suitable for a certain site and application.

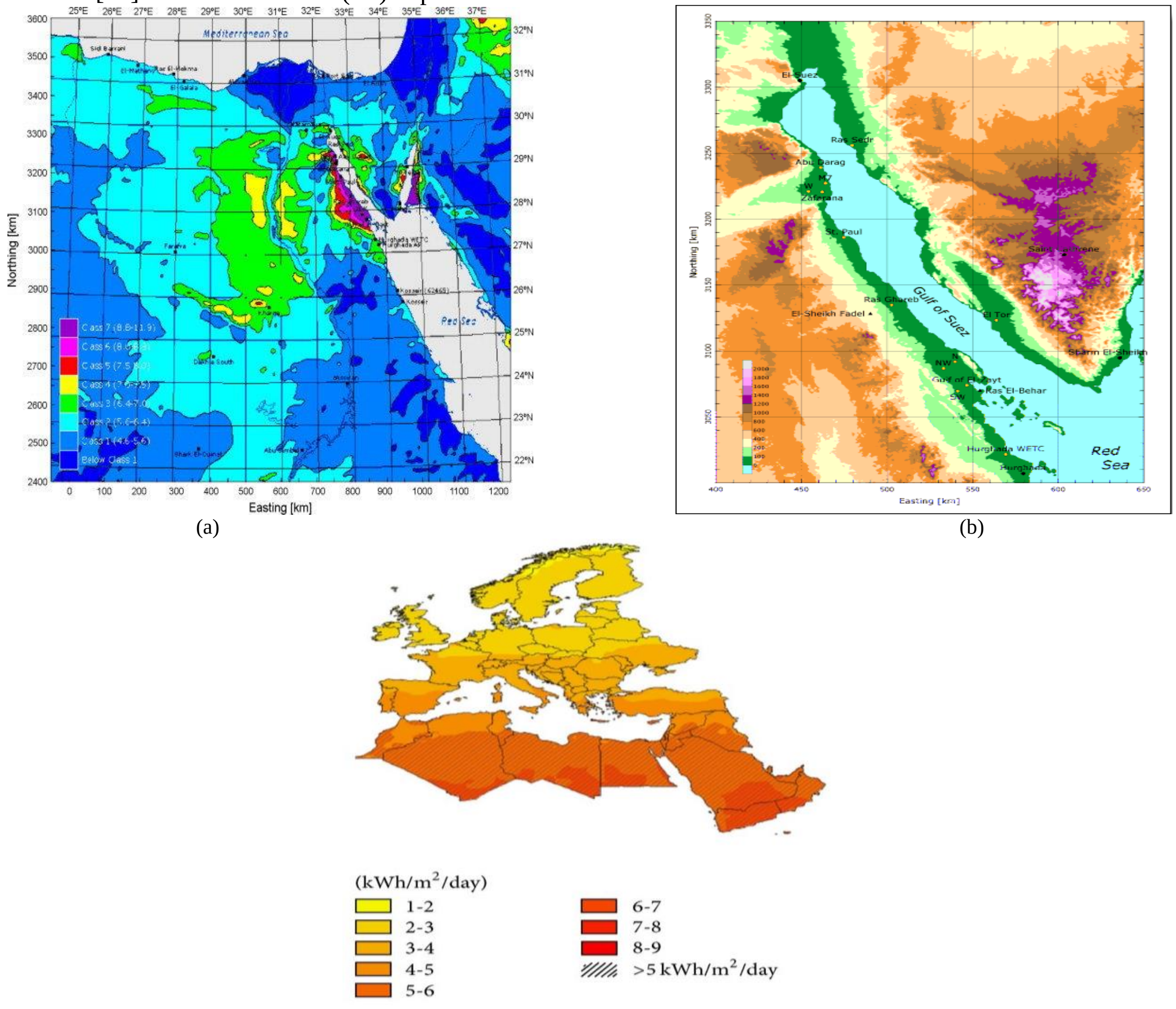
Because of the inherent fluctuations in solar and wind energy resources, the independent use of an individual power source usually results in a very large generation and storage system, which in turn requires a higher operating and life cycle cost [5, 6]. Therefore, the hybrid solar-wind system is usually adopted, which can leverage the strengths of each technology to provide a more reliable and less costly power supply in remote areas [3, 4]. Energy storage systems such as batteries, fuel cells, flywheels, supercapacitors, molten salt, compressed air and hydroelectric pumped storage (HPS) can be a suitable solution to overcome the problems associated with the irregular nature of renewable sources.

Hydrogen tanks as a mean of energy storage in renewable energy systems has been proposed in many research studies [7-9]. Hydrogen storage performs well in both long and short-term purposes and in some cases for enhancement of the dynamic performance of the system, super capacitors are advisable to be integrated [10]. In comparison with diesel generator as a method of energy storage, hydrogen-based energy storage systems do not need any external supply of fuel and no greenhouse emissions are exhausted into the atmosphere. In addition, due to the continuous increase in the fuel cost and the observable reduction in the cost of FC, it is expected that the hydrogen storage systems will be economically feasible for application as a method of electric energy storage in the form of hydrogen gas [11].

In hybrid renewable energy systems based on hydrogen storage, during the hours of excess energy from the PV and wind systems, electrolyzer coverts the surplus electrical energy into hydrogen stored in the hydrogen tanks. When the peak demand for electricity occurs and the wind speed and solar radiation are not sufficient to satisfy the load demand by converting the stored hydrogen into electricity through the FC [12].

Due to the fact that hybrid PV/wind power systems depend mainly on sources having intermittent characteristic (solar radiation and wind speed), it is a great challenge to design such a system with an acceptable reliability when considering the investment and operating costs of each component in the system. Therefore, the main goal is the optimal configuration of an economic and reliable power supply. In the literature, there are many research papers, which offer different methods and algorithms for optimal design of hybrid PV/wind power systems [13-28].

In [13, 14] the optimal sizing of a PV/wind/diesel hybrid system was presented using Strength Pareto evolutionary algorithm by formulating two objective functions, minimization of system cost and greenhouse gases emissions. In [15-17], Genetic algorithm (GA) was used for optimal sizing and configuration of a hybrid PV/wind power system with battery storage under different objective functions; the reliability of the system under weather conditions variations, minimizing the annual cost of the system and to minimize the loss of power supply probability (LPSP). Particle Swarm Optimization (PSO) has been used in many research studies for optimal sizing of hybrid renewable energy systems [18-20]. Simulated Annealing (SA) optimization strategy was used for optimal sizing of hybrid PV/wind energy conversion system, while the objective function was to minimize the total energy cost of the hybrid system [21]. The response surface methodology (RSM) was used for optimizing the size of an autonomous PV/wind system with energy storage in some studies [22]. The results obtained from RSM optimization were confirmed using autonomy analysis and loss of load probability, then the results of this work were used in [21] to be compared with the results obtained from simulated annealing optimization. In [23] Pattern Search (PS) optimizer and sequential Monte Carlo Simulation (SMCS) are combined to obtain the minimum total cost of the system and satisfy the reliability requirements from the consumer side. A comparison with a hybrid GA-SMCS was also performed, from which the PS-SMCS gave a better performance. In [24] cuckoo search optimization algorithm has been used for optimal sizing of an isolated PV/wind/diesel/ battery energy system, while the proposed technique provided high accuracy when compared with GA and PSO. Multi-Objective Self-Adaptive Differential Evolution (MOSaDE) algorithm has been used for optimal sizing and operation of a hybrid PV/wind/diesel microgrid system with battery storage for the city of Yanbu, Saudi Arabia, while, the multi-objective optimization approach is used to reduce the computational time [25]. Optimal sizing and placement of a grid-connected PV-wind-battery storage microgrid using Artificial Bee Colony optimization technique, while the IEEE 30-bus system was used for the application of the optimal operation [26]. In [27, 28] the Hybrid Optimization of Multiple Energy Resources (HOMER) software has been used for optimum sizing of hybrid wind/PV/diesel system in Malaysia in which the weather conditions, maximum availability and minimum cost were considered respectively.



(a)

(b)

(c)

Fig. 1. (a) Wind Resource Map of Egypt: mean Wind Speed at 50 m a.g.l. Determined by Mesoscale Modelling (Wind Atlas for Egypt, 2006), (b) Wind Atlas of Gulf of Suez, Egypt. (c) Egypt's Solar Potential.

This paper investigates on the economical design of a grid-dependent hybrid PV/wind power system with electrolyzer, hydrogen tank and fuel cell. In this way, a MFFA has been exploited for minimizing the cost of energy (COE) generated from this system over its 25 years lifetime and subjected to reliability constraints in the form of LPSP. This study is in line with the plans of the Egyptian government to encourage Egyptian and foreign companies to expand for establish renewable energy projects where it has developed catalyst laws and legislations in this field [37]. The current sites of renewable energy projects in Egypt, mostly concentrated in the Gulf of Suez and south Egypt because of its high wind speed and high solar radiation [37]. Moreover, the Solar and wind speed spectrum have been shown in Fig. 1. The figure shows that the Gulf of Suez area is very large and promised to implement wind energy power plant. Also, the figure shows this region is suitable for implementation of the PV power plant. Moreover, this region has three wind power plant in Gabal Elzeit with 380 MW total power and a grid connected PV project around 20 MW, in Hurghada [37]. Therefore, the area of Ataka at the Gulf of Suez has been chosen to be the place of study where it lies on the shore of the Red sea (latitude 30.0, longitude 32.5) [29]. The system cost includes the annual interest of capital investment cost, operation and maintenance, replacement cost, cost of energy sold and purchased from the external grid as well as the penalty cost due to unacceptable reliability and fluctuation rate in the energy exchange with the utility grid.

The main contribution of this paper can be written as follows:

- A MFFA has been presented; this modification is proposed to reduce the processing time of the recent FFA.
- The MFFA algorithm has been applied for design the Hybrid energy system. Moreover, the results have been analyzed and compered with these of the conventional one.
- The objective function consists of two terms; the cost of energy and LPSP.
- A real case of study has been selected in Egypt to validate the presented optimization technique.

The paper is organized as follows: Section 2 briefly describes the subsystems of the hybrid system and their corresponding models. The operation strategy is demonstrated in Section 3. Optimization problem statement, FFA and MFFA are discussed in Section 4. Simulation results and discussions are summarized in Section 5. Finally, Section 6 is devoted to conclusion.

II. Construction Description of the Proposed System

A simplified single line diagram of the proposed grid-connected hybrid system is shown in Fig. 2. The proposed system includes seven main components in addition to the AC and DC buses. The system includes wind turbine generators, photovoltaic arrays, electrolyzer, hydrogen tanks, fuel cells, converter and the electric utility grid. During the hours of low demand of electric energy and when the generated power from the renewable sources exceeds the load demand the excess energy is supplied to the electrolyzer to store the energy in the form of chemical energy (hydrogen) in hydrogen tanks. If the redundant energy is higher than that required to fully charge the hydrogen tanks, then the surplus power is supplied to the grid. On the other hand, during the nighttime, in the time of maximum load, when the generation is less than the load demand, the energy stored in the hydrogen tanks is applied to the fuel cell to convert it again to electric energy serving the load. In the case of insufficient supply from the renewable sources, the energy deficit is covered by the electric grid. The models of the previously mentioned components are described in the following sections.

A. PV System

The output power generated from a PV module in terms of the solar radiation and the ambient temperature can be expressed as [29]:

$$P_{PV}(\mathrm{t}) = n_{PV} P_{PV\_rated} \eta_{PV} \eta_{Wire} \frac{G(t)}{G_{nom}} (1 - \beta_T (\mathrm{T}_C(t) - \mathrm{T}_{nom})) \tag{1}$$

where, n_{PV} is the number PV modules, P_{PV_rated} is the rated power of the PV module at standard operating conditions (G_{nom} = 1000W/m^2 and T_{nom} = 25ºC), η_{PV} is conversion efficiency of the PV module, and η_{wire} is the wire efficiency. $G(t)$ is the ambient solar radiation intensity, G_{nom} is the intensity of solar radiation under standard conditions, β_T is the temperature coefficient of power of the selected PV module, $T_c(t)$ is the cell temperature, and T_{nom} is the cell temperature under standard conditions of operation, respectively.

The efficiency of the PV modules inherent includes the efficiency of the maximum power point tracking (MPPT) system. As stated in [30], using the MPPT system will increase the energy generated from the photovoltaic modules by about 30%, it is economically feasible to incorporate them in such hybrid systems. As a result, in the system under study, it is supposed that the PV modules are fixed on tracking system having an efficiency of 95%.

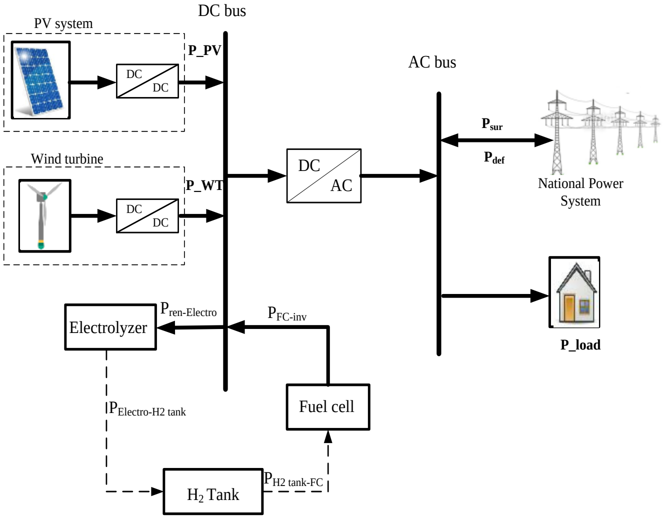


Fig. 2. Single Line Diagram of the Grid-Dependent Proposed Hybrid System.

B. *Wind Turbine*

Depending on the fundamentals of wind energy, the expected energy supplied by a wind turbine can be described as follow [29]:

$$P_{WT}=\begin{cases}0 & u(t)<u_{cut-in}\\ \dfrac{n_{wind}\eta_{wind}P_{R\_WT}(u^2(t)-u^2_{cut-in})}{(\mathrm{u}^2_{rated}-u^2_{cut-in})} & u_{cut-in}<u(t)<u_{rated}\\ n_{wind}\eta_{wind}P_{R\_WT} & u_{rated}<u(t)<u_{cut-off}\\ 0 & u(t)>u_{cut-off}\end{cases} \tag{2}$$

where, P_{WT} is the actual power generated from wind turbine, P_{R_WT} is the wind turbine rated power, n_{wind} is the number of wind turbine, η_{wind} is the efficiency of the wind system. $u(t)$ is the wind speed at time t, u_{cut-in} is the cut-in wind speed, at which the turbine start operation , u_{rated} is the wind speed at rated power, and $u_{cut-off}$ is the cut-off wind speed, after which the wind turbine must be shut down for safety reasons. In this study, Bergey Wind Power's BWC Excel-R/48 wind turbine, which has a rated power of 7.5kW, which provides 48V DC is proposed. The detailed model of the wind turbine suggested in this study is described in [11]. Considering the fact that the wind speed changes with height, the wind speed at a desired hub height as a function of the wind speed measured at the anemometer height is given as follow [30]:

$$\frac{u_2}{u_1}=\left(\frac{h_2}{h_1}\right)^{\alpha_{WT}} \tag{3}$$

where, u_2 is the wind speed at the wind turbine hub with height h_2, u_1 is the wind speed at the reference point h_1, and α_{WT} is the friction coefficient. According to the recommendations of IEC standards, the value of coefficient of friction is taken as 0.11 for extreme wind conditions, and 0.20 for normal wind conditions. The renewable power injected to the DC busbar is the sum of the power generated from the PV panels and the wind turbine generators. The renewable power takes two different streams. The first part ($P_{ren-inv}$) flows through the DC/AC inverter to supply the load connected at the AC bus and the surplus power fed into the grid. The second stream ($P_{ren-Electro}$) is used by the electrolyzer for producing hydrogen.

C. *Electrolyzer*

The electrolyzer uses the process of water electrolysis to decompose the water to its original molecules, hydrogen and oxygen, using a DC current flows between two electrodes. Then the hydrogen is collected around the surface of the anode. As stated in [31] in most water electrolyzer, the produced hydrogen is collected at a pressure of 30 bar, which is higher than the reactant pressure of the hydrogen supplied to the Proton Exchange Membrane Fuel Cell (PEMFC), which is around 1.2 bar. The hydrogen produced from the electrolyzer might be directly applied to the hydrogen tank as in most studies, or the pressure of the produced hydrogen is raised to about 200 bar using a compressor in order to raise the energy stored density [31]. In other studies, the hydrogen obtained from the electrolyzer is supplied to a low pressure tank, then when the tank is fully charged, a compressor is used to pump the stored hydrogen to a high pressure tank. Thus, the energy consumed by the compressor is reduced as it is not in operation all time [12]. In this study, the hydrogen produced is directly injected to the hydrogen tank. The electrolyzer is modelled by means of the power transferred from the dc bus-bar to the hydrogen tank and is described as follows [11, 33]:

$$P_{Electro-\tan k}=P_{ren-Electro}\times\eta_{Electro} \tag{4}$$

where, $P_{Electro-tank}$ is the output power of the electrolyzer, which is supplied into the hydrogen tank, $P_{ren-Electro}$ is the input power to the electrolyzer, and η_{Electro} is the efficiency of the electrolyzer, which is supposed to have a constant value for the whole simulation time.

D. *Hydrogen Tank*

In this paper, the hydrogen tank is modelled by means of the amount of energy stored in the hydrogen tank and at any time of simulation *t* is mathematically calculated from the following equation [33]:

$$E_{\tan k}(t)=E_{\tan k}(t-1)+\left(P_{Electro-\tan k}(t)-\frac{P_{\tan k-FC}(t)}{\eta_{storage}}\right)\times\Delta t \tag{5}$$

where, $E_{tank}(t)$ is the energy stored in the hydrogen tank at any time t, $E_{tank}(t-1)$ is the energy stored in the tank at time (t-1), $P_{tank-FC}(t)$ is the equivalent power drawn from the hydrogen tank and injected into the fuel cell, and $\eta_{storage}$ is the efficiency of the storage tank and it represents the losses due to self-discharge of the tank itself and it is taken as 95% for all operating scenarios [32]. Δ*t* is the time interval in the simulation process, which is considered equal to an hour in this paper. At any simulation time *t* the mass of the hydrogen stored in the tank is calculated as follow [33]:

$$M_{\tan k}(t)=\frac{E_{\tan k}(t)}{HHV_{H_2}} \tag{6}$$

where, HHV_{H2} is the higher heating value (HHV) of the hydrogen. HHV of a certain fuel is defined as the amount of heat produced by a specific amount at standard temperature (25°C), when it is combusted and the products of the combustion process are returned again to 25°C. According to [34], the value of HHV for hydrogen is taken as 39.7kWh/m^2. The stored energy in the tank at any time t occurs between a predefined upper and lower limits. The upper limits are described by the maximum capacity of the tank. For some problems associated with the nature of the hydrogen itself, it is recommended that a small amount of the hydrogen stored is not discharged, which forms the lower limit of the hydrogen tank (here 5%). Therefore,

$$M_{\tan k,\min}\le M_{\tan k}(t)\le M_{\tan k,\max} \tag{7}$$

where, $M_{tank,min}$ is the lower limit of the hydrogen tank, $M_{tank}(t)$ is the mass of hydrogen stored in the tank at any time t, and $M_{tank,max}$ is the upper limit of the hydrogen tank.

E. Fuel Cell

The basic operation of the hydrogen fuel cell is extremely simple. In the hydrogen FC the electrolysis is being reversed – the hydrogen and oxygen are recombining, and an electric current is being produced. During this chemical reaction, electric energy is produced in the form of electrons, which are released in the process of ionization of the hydrogen. PEMFC is commercially produced in high generating capacities and reliable and good dynamic response thanks to its short power release time of about 1-3 s [31]. In this study the efficiency of the fuel cell is assumed to be constant and taken as 50%, then its output can be simply calculated as a function of the input power and efficiency. Therefore,

$$P_{FC-inv} = P_{\tan k-FC} \times \eta_{FC} \tag{8}$$

where, $P_{tank\text{-}FC}$ is the input power to the fuel cell, which describes the mass of hydrogen consumed in the chemical reaction, and η_{FC} is the efficiency of the fuel cell.

F. DC/AC Converter

The inverter is used to convert the DC power from the renewable sources and the fuel cell into the form of AC power to cover the load demand and the excess power is supplied to grid. According to [11], the efficiency of the inverter (η_{inv}) is assumed to be constant at 90% for the whole simulation time. Therefore, the power output from the inverter is described as follows:

$$P_{inv-AC} = \left(P_{FC-inv} + P_{ren-inv}\right) \times \eta_{inv} \tag{9}$$

where, $P_{FC\text{-}inv}$ is the output power from the fuel cell, and $P_{ren\text{-}inv}$ is the power generated from the renewable sources and directly supplied to the load.

III. Operation Strategy

The operation of the proposed hybrid renewable system is summarized in the following operation scenarios.

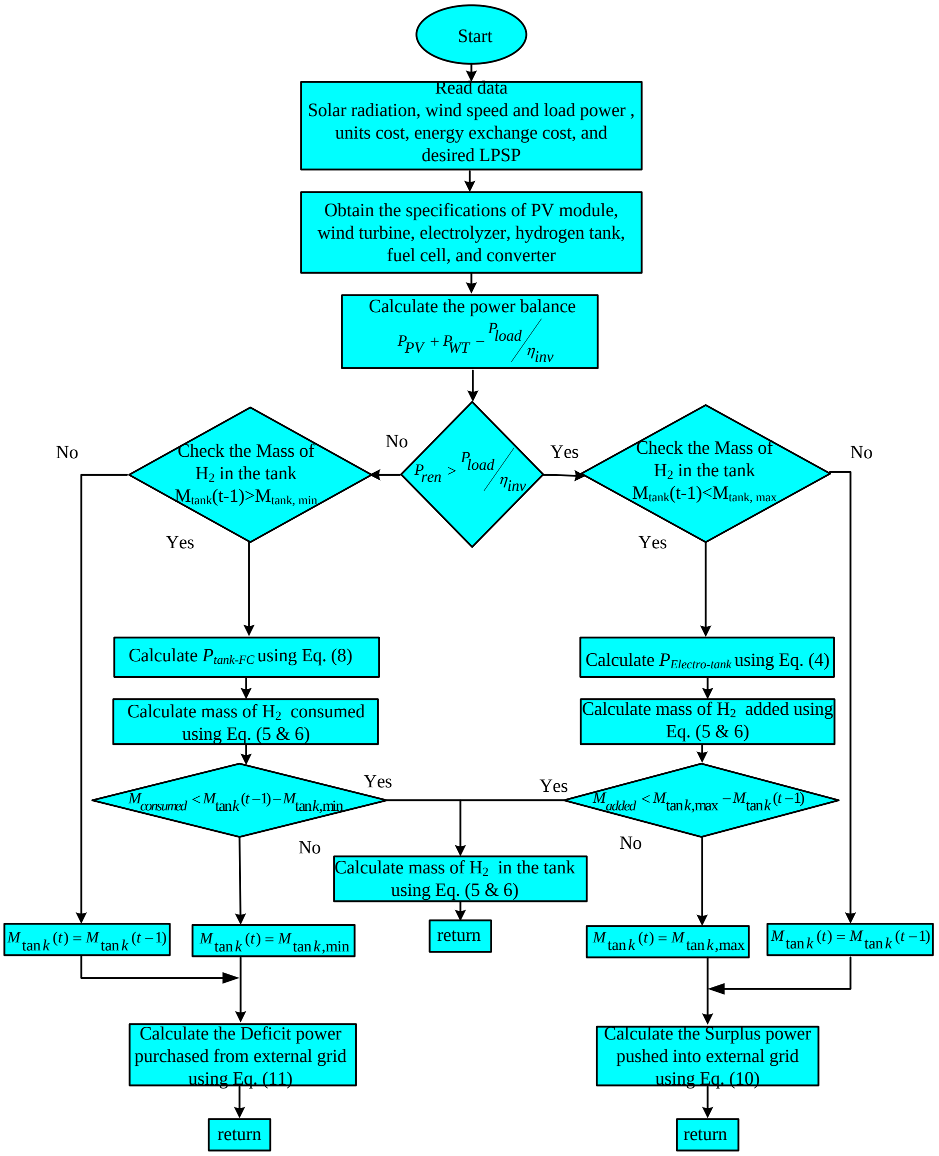

Start
Read data
Solar radiation, wind speed and load power , units cost, energy exchange cost, and desired LPSP
Obtain the specifications of PV module, wind turbine, electrolyzer, hydrogen tank, fuel cell, and converter
Calculate the power balance
$P_{PV} + P_{WT} - P_{load}/\eta_{inv}$
$P_{ren} > P_{load}/\eta_{inv}$
No
Yes
Check the Mass of H_2 in the tank $M_{tank}(t-1) > M_{tank, min}$
Check the Mass of H_2 in the tank $M_{tank}(t-1) < M_{tank, max}$
No
Yes
Yes
No
Calculate $P_{tank\text{-}FC}$ using Eq. (8)
Calculate $P_{Electro\text{-}tank}$ using Eq. (4)
Calculate mass of H_2 consumed using Eq. (5 & 6)
Calculate mass of H_2 added using Eq. (5 & 6)
$M_{consumed} < M_{\tan k}(t-1) - M_{\tan k,min}$
Yes
Yes
$M_{added} < M_{\tan k,max} - M_{\tan k}(t-1)$
No
No
Calculate mass of H_2 in the tank using Eq. (5 & 6)
return
$M_{\tan k}(t) = M_{\tan k}(t-1)$
$M_{\tan k}(t) = M_{\tan k,min}$
$M_{\tan k}(t) = M_{\tan k,max}$
$M_{\tan k}(t) = M_{\tan k}(t-1)$
Calculate the Deficit power purchased from external grid using Eq. (11)
Calculate the Surplus power pushed into external grid using Eq. (10)
return
return

Fig. 3. Flowchart Describing the Operation of the Hybrid System.

If the generated power from the renewable sources ($P_{ren}(t) = P_{PV}(t) + P_{WT}(t)$) equal to the load power ($P_{load}(t)$), $P_{ren}(t) = P_{load}(t)/\eta_{inv}$, then the whole renewable power is supplied to the load.

If $P_{ren}(t) > P_{load}(t)/\eta_{inv}$, the excess power is applied to the electrolyzer. If the difference exceeds the rated power of the electrolyzer, then the surplus power ($P_{sur}(t)$) is injected into the electric grid. Therefore,

$$P_{sur}(t) = P_{ren-inv}(t) \times \eta_{inv} - P_{load}(t) \tag{10}$$

If $P_{ren}(t) < P_{load}(t)/\eta_{inv}$, then the shortage in load demand is covered by the fuel cell. If the power needed to cover the load exceeds the rated of the fuel cell, then the electric grid supplies the power deficit ($P_{def(t)}$). Therefore,

$$P_{def}(t) = P_{load}(t) - \left(P_{ren-inv}(t)+P_{FC-inv}(t)\right) \times \eta_{inv} \tag{11}$$

A flowchart describing the operation of the proposed hybrid system is shown in Fig. 3.

IV. Optimization Problem

A. Cost of Energy Generated from the Proposed Hybrid System

The annual interest of capital cost for each component in the system as apart from the initial investment cost (C_{cap_i}) for each component *i* is calculated as follows [24]:

$$C_{ann\_cap\_i} = C_{cap\_i} * CRF(r, M_i) \tag{12}$$

where, *i* refers to each component in the system including; wind turbine, photovoltaic system, electrolyzer, hydrogen tank, fuel cell, and DC/AC converter. *CRF* is the capital recovery factor, *r* is the rate of interest (here = 0.06), and M_i is the lifetime of each sub-system [24]. CRF is calculated from the following formula [24]:

$$CRF(r, M_i) = \frac{r(1+r)^{M_i}}{(1+r)^{M_i} - 1} \tag{13}$$

The operating and maintenance cost ($C_{o\&m}$) of the hybrid renewable energy system is the major running cost of the system as there is no fuel cost. Starting from the fact that each component in the power system has its own lifetime, which is different from the lifetime of the whole system as an economic project (M_{sys} = 25 years), so that there is a need for replacement for the individual subsystems. The annual cost for replacement (C_{rep_ann}) for each subsystem i is calculated as follows [35]:

$$C_{ann\_rep} = C_{rep\_i}\,\frac{\left(M_{sys} - M_i\right)}{M_i} \tag{14}$$

In the case of a grid connected power system, the fluctuation of the injected energy into the grid and LPSP causes a penalty cost if they exceed the predefined values. In this paper, LPSP is limited to not exceed the predefined value (βL = 0.05) [35] and it is calculated according to the following formula [35]:

$$LPSP = \frac{\sum_{i=1}^{8760}[P_{load}(t_i) - (P_{PV}(t_i) + P_{WT}(t_i) + P_{FC-inv}(t_i))]}{\sum_{i=1}^{8760} P_{load}(t_i)} \tag{15}$$

According to [35] the maximum power fluctuation rate should not exceed 33% of the installed power of the system for an interval of 10 minutes. The power fluctuation rate (D_{gs}) is calculated as follows [35]:

$$D_{gs} = \frac{P_{sur\_\max} - P_{sur\_\min}}{\Delta t} \tag{16}$$

where, P_{sur_max} is the maximum value of the power supplied to grid and P_{sur_min} is the minimum value of the surplus power. During operation the fluctuation rate is limited to the predefined value β_g. If the value of *LPSP* and D_{gs} exceed their suggested values, then the penalty cost (C_{pc}) of the system is calculated according to the following formula [35]:

$$\begin{aligned} C_{pc} &= C_{pc\_1} \times (\mathrm{LPSP} - \beta_L) \times \sum_{i=1}^{N} P_{load}(t_i) \\ &+ C_{pc\_2} \times \frac{D_{gs} - \beta_g}{\beta_g} * 100 \end{aligned} \tag{17}$$

Where, C_{pc_1} is the penalty cost for the shortage of supply and C_{pc_2} are the penalty cost for fluctuation in the supply. During simulation and the operation of the optimal sizing model the penalty costs are considered as C_{pc_1} = 100$/kWh and C_{pc_2} = 50000$/%.

To make a full use of the PV and wind complementary characteristics, the relative fluctuation rate is adopted for the renewable generation with respect to the load power [35]:

$$D_{load} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_{PV}(t_i) + P_{WT}(t_i) - P_{load}(t_i)\right)^2}}{\overline{P}_{load}} \tag{18}$$

where, $\overline{P}_{load}$ is the average load power. Smaller value of D_{load} means that the power generation is closer to the load demand and ensures better utilization of the sources. The hybrid system will ensure a full utilization of renewable sources characteristics, if the relative fluctuation is smaller than the allowable reference (ε_L = 2) [35].

When the power generated from the renewable sources is greater that the load demand and the power consumed by the electrolyzer, the surplus is sold to the grid (C_{gs}). In opposite when the renewable generation and the power supplied by the fuel cell do not cover the load demand, the deficit is purchased from the grid (C_{gp}). The cost of energy sold and purchased is calculated as follows:

$$C_{gp} = N_{gp} \times C_p \tag{19}$$

$$C_{gs} = N_{gs} \times C_s \tag{20}$$

where, C_p (here 0.08$/kWh) the cost of kWh purchased from the external grid, C_s (here 0.2$/kWh) is the cost of kWh supplied to the grid and N_{gp} is the total units purchased from the grid, and N_{gs} is total energy units supplied to the grid. The specific characteristics of components applied in system simulation are depicted in Table I [11]. The annual cost of the hybrid renewable system (C_{ann_tot}) is calculated as the sum of the individual costs for the whole system [35]:

$$C_{ann\_tot} = C_{ann\_cap} + C_{ann\_rep} + C_{\text{O\&M}} + C_{pc} + C_{gp} - C_{gs} \qquad (21)$$

The net present cost (NPC) of the system is calculated as follow [24]:

$$NPC = \frac{C_{ann\_tot}}{CRF} \qquad (22)$$

The cost of energy (COE) from the hybrid system in ($/kWh) and formulated in the following equation [24, 35]:

$$COE = \frac{C_{ann\_tot}}{\sum_{h=1}^{h=8760} P_{load}} = \frac{NPC}{\sum_{h=1}^{h=8760} P_{load}} \times CRF \qquad (23)$$

B. Objective Function and Constraints

In this paper, as introduced a Minimizing the objective function of COE, LPSP for the optimization algorithm as a function in Wind turbine, PV array, Electrolyzer, hydrogen tank, FC and DC/AC converter can be expressed as:

$$\min f = \min(obj\_func) \qquad (24)$$

where;

$$obj\_func = \lambda_1 * COE + \lambda_2 * LPSP$$

λ_1, λ_2 are chosen as trial and error to reach the best results. In this work, the values of λ_1, and λ_2 are 0.5, and 0.5 respectively.

MFFA optimization algorithm has been used for optimal sizing of the proposed hybrid system under the following constraints:

$$\begin{aligned} LPSP &\le \beta_L \\ M_{\tan k,\min} \le M_{\tan k}(t) &\le M_{\tan k,\max} \\ D_{gs} &\le \beta_g \\ D_{load} &\le \varepsilon_L \end{aligned} \qquad (25)$$

C. Farmland Fertility Algorithm

In this paper, the recent optimization algorithm of farmland fertility, which is presented for the first time in [36], is applied for design the renewable energy resources system. The farmer in their land bases the idea of FFA optimization algorithm on the determination of the quality the soil of each region of farmland. The main reason of variations of Soil's quality is based on consisting and adding of special materials. Therefore, farmers use different materials to improve the quality of farmland. In other words, these materials when added to the soil in farmland, they may improve or reduce the quality of the soil. The determination of these materials to improve the soil quality is based on trying materials based on the farmers experience with each type of soil and the previous results of improving in the soil quality for each time. There are main steps for determine the best quality and best materials for improving the soil quality, which can be mathematically described as follows:

1) First stage initialization: At this stage, the generation of the initialized population is based on the following criteria:

- Taking the number of sections of farmland in to consideration (number of parts for optimization problem and is expressed by k).
- Considering the number of available solutions in each section, (the number of existing solutions in each section of farmland and is expressed by n).

The following mathematical equation is to describe this first stage to generate the total number of population N:

$$N = k * n \qquad (26)$$

where, *k* is an integer number and greater than zero. The constant k is considered for partition of search space (As farmers divide their land into different parts). Its value is between 1 and *N*. In this paper, it is determined by trial and error to reach the best solution and its value equals to 2. In [36], the authors refer to if the value is greater than 8, the FFA algorithm gives very poor performance. So, k may be considered as $2 \le k \le 8$. Moreover, n is an integer number. It is also determined by trial and error in this work. It describes the available solutions from each search space. The following equation generates the random solution in the first stage taking into consideration the upper and lower limits of the each variable Uj and Lj, respectively.

$$X_{ij} = L_j + rand(0,1) \times (U_j - L_j) \qquad (27)$$

TABLE I. SPECIFICATIONS OF THE SYSTEM COMPONENTS [11]

Component	Capital cost (US$/unit)	Replacment cost (US$/unit)	O&M (US$/ unit-yr)	Lifetime (yr)	Efficiency (%)	Unit
Wind turbine	19400	15000	75	20	-	7.5kW
PV array	7000	6000	20	20	-	1kW
Electrolyzer	2000	1500	25	20	75	1kW
Hydrogen tank	1300	1200	15	20	95	1kg
Fuel cell	3000	2500	175	5	50	1kW
DC/AC converter	800	750	8	15	90	1kW

where, j=[1. . ..D] represents dimension of the variables x in the optimization problem and is based on the total number of population and is equal to [1. . ..N].

In this paper, for the presented problem the x represents the size of RES plants which are (number of PV units, number of wind turbines, rated power of the electrolyzer, maximum capacity of the hydrogen tank, rated power of the fuel cell, and rated power of the DC/AC converter) which is required to determine the best optimal size RES plants.

Fig. 4 shows that the farmland is allocated into 3 sections, each section has its own local memory and a global memory and part A has soil with lowest quality. In the following part, the mathematical expression of each stage will be expressed.

2) Second stage: determination of the quality of soil in each section of farmland: In this stage, the fitness of all the existing solutions in the search area have been evaluated. Moreover, the soil quality of each of the parts of the farmland (sections of the search space) can be determined as following:

$$Section_s = X(aj), \;\; a = n^*(s-1) : n*ss = \{1, 2, ...k\},$$
$$j = \{1, 2, ...4\} \tag{28}$$

The average value of the existing solutions in each part of the farmland is used to determine the quality of each section.

$$Section_s = X(aj), \;\; a = n^*(s-1) : n*ss = \{1, 2, ...k\},$$
$$j = \{1, 2, ...4\} \tag{29}$$

At this stage of the farmland fertility the individual sections of the farmland have been determined. In addition, the solutions and their average for each section have been calculated.

3) Third stage update memories: In this stage, the local memory and global memory have been updated. For1 each section of the farmland, some of best solutions are stored in the local memory of that section and best solutions of these individual sections are stored in a global memory. The number of best local memory and the number of best global memory are determined according to Eqs. (30) and (31), respectively.

$$M_{local} = round(t*n), \qquad 0.1 < t < 1 \tag{30}$$

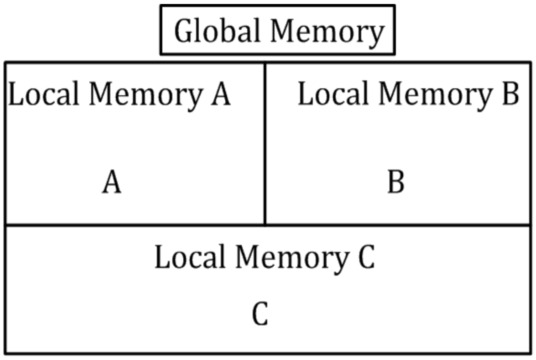


Fig. 4. Partitioned Example of Farmland and Local Memory and Global Memory.

$$M_{Global} = round(t*n), \qquad 0.1 < t < 1 \tag{31}$$

where, the M_{Global} express the number of stored solutions in global memory. While the number of the stored solution in local memory is M_{local}. Based on the fitness and suitability of the local and global memories, the solutions have been placed. Moreover, at this stage both memories are updated.

4) Fourth stage: changing quality of the soil in each section of farmland: In the third stage, the soil quality of each section has been determined using equation (28). The best solutions of each part of the farmland has been stored in the corresponding local memory. Moreover, the best solution for all sections have been stored in the global memory. In this stage, the solutions of the worst section should be updated and more changes should be happened to improve its quality. So the all existing solutions in worst section can be combined with the best solution of the global memory. This can be mathematically expressed as in Eqs. (32) and (33):

$$h = \alpha * rand(-1, 1), \tag{32}$$

$$X_{new} = h*(X_{ij} - X_{M_{Global}}) + X_{ij} \tag{33}$$

where, $X_{MGlobal}$ is a random solution among existing solutions in the global memory and α is a number between zero (0) and one (1) that should be valued at the beginning of the farmland fertility. X_{ij} is a solution in worst part of farmland that is selected to apply changes. Moreover, h is a decimal number which is computed from Eq. (32). Consequently, X_{new} is the new solution. This solution has been take place to change.

After applying the previous change: For updating the solutions of other sections; a combination between the existing solutions of each section (apart from the worst section) are combined as the following equations (34) and (35):

$$h = \beta * rand(0, 1), \tag{34}$$

$$X_{new} = h*(X_{ij} - X_{uj}) + X_{ij} \tag{35}$$

where, X_{uj} is a random solution among existing solutions in the search area. This means that, a random solution is selected from all the existing solutions in sections. β is a number in the range between zero (0) and one (1), which should be determined at the starting of the farmland fertility process. X_{ij} is a solution belongs to the worst section and has been chosen for applying changes in the solution. *h* is a decimal number obtained from Eq. (34). After applying changes a new solution X_{new} is delivered.

5) Fifth stage: soil's combination: At this stage, as we theoretically convey algorithm in part (1), depending on the best solutions available in the local memory ($Best_{Local}$) of each section in the final stage, farmers decide to merge each soil within the parts of farmland. Therefore, there is a provision about the combination with the best in local memory. So that, not all available solutions are combined with local memory in all sections and at this stage, in order to improve the quality of solutions in each part of farmland, some of the available solutions in all locations are combined with the best solution ever found (*Best_{Global}*). The combination of considered solution with $Best_{Global}$ or $Best_{Local}$ is determined by Eq. (36).

$$H = \begin{cases} X_{new} = X_{ij} + \omega_1 * (X_{ij} - Best_{Global}(b)) & Q > rand \\ X_{new} = X_{ij} + rand(0,1) * (X_{ij} - Best_{local}(b)) & else \end{cases} \tag{36}$$

In Eq. (36), a new solution may be created by two different methods. In this equation, Q is a parameter between zero (0) and one (1) and has to be adjusted at the start of the optimization algorithm. The Q parameter determines to what degree the solutions are combined with best global solution ($Best_{Global}$). ω_1 is an integer number and should be evaluated at the start of the process and according to repetition of optimization algorithm its amount has been gradually decreased (Eq. (37)). X_{ij} is a solution that to apply the changes is selected from all sections. As a result, according to the applied changes, a new solution X_{new} has been produced.

$$\omega_1 = \omega_1 * R_v, 0 < R_v < 1 \tag{37}$$

6) Sixth stage: final conditions: At this stage, the existing available solutions in search area are evaluated according to the objective function. Whatever the number of sections is, this stage is performed on all existing solutions in the search field. Thus, the fitness the degree of suitability of each of the available solutions is determined in the search space. At the end of the farmland fertility, are investigated final conditions. If we confirm final condition, the algorithm ends. Otherwise, the algorithm will continue its work to establish final conditions. The flowchart of the FFA optimization algorithm has been shown in Fig. 5.

D. Modified FFA Algorithm

In this paper a modified FFA algorithm has been proposed and tested. Moreover, the modified FFA algorithm is compared with the conventional FFA algorithm which presented in [36]. The modified FFA algorithm is to combine the solutions of the fourth and fifth stages to reduce the processing time of the execution of the optimization algorithm.

A new random number m has been created. This random number is between 0 and 1. This number reduces the implementation time of the algorithm. While, the FFA algorithm tries three solutions equations (33, 35 and 35) in a cascade manner which increase the processing time of the algorithm. Moreover, this random number is to determine the balancing between the local and global optimization solutions.

In the proposed FFA algorithm, the new solution is defined as a combination of equations (32)-(36) for updating the new solution best global solution in the previous iteration and random solution based on a random solution of the best solution in the local memories:

$$H = \begin{cases} X_{new} = X_{ij} + \omega_1 * (X_{ij} - Best_{Global}(b)) & Q > rand \\ X_{new} = X_{ij} + rand(0,1) * (X_{ij} - Best_{local}(b)) + h * (X_{ij} - X_{MGlobal}) & else \end{cases} \tag{38}$$

where, X_{ij} is a solution, which is considered to update for each section. $Best_{Global}$ is the best solution in the global memory. $Best_{Local}$ is the best solution in the local memory. $X_{MGlobal}$ is a random solution among existing solutions in the global memory. *h* is defined in equation 33. ω_1 as a parameter of the farmland fertility and is updated as follows:

$$\omega_1 = \omega_1 * R_v, \quad 0 < R_v < 1 \tag{39}$$

The flowchart of the proposed MFFA is shown in Fig. 6.

The optimization problem is to minimize the objective functions of equation (23) and determine the best optimal size of RES plants which are (number of PV units, number of wind turbines, rated power of the electrolyzer, maximum capacity of the hydrogen tank, rated power of the fuel cell and rated power of the DC/AC converter). In this paper, each of the proposed optimization algorithms operates according to the following steps:

1) Run the program of designing the RES based on the energy balance.

- Generation of initial population, Equation (27)
- Run the program of designing the hybrid energy system described by equations from (1) to (23) and illustrated in the flowchart of Fig. 3.
- Evaluation process of the fitness function for all agents/positions. Equation (24)

2) Run the optimization algorithm with the RES based on equations (28) and (35) as the following processes:

- Updating the position and sizing the RES elements, according the nature of each optimization algorithm.
- Run the program of designing the RES. Equations from (1) to (23) and illustrated in the flowchart of Fig. 3.
- Evaluation process of the fitness function for all agents/positions, Equation (24).
- Check if the proposed system meets end criterion, If “No”, repeat the previous three processes. If “Yes”, the program will be stopped and will go to the next step.
- Type the results such as sizing of system components and the best optimum values of COE and LPSP.

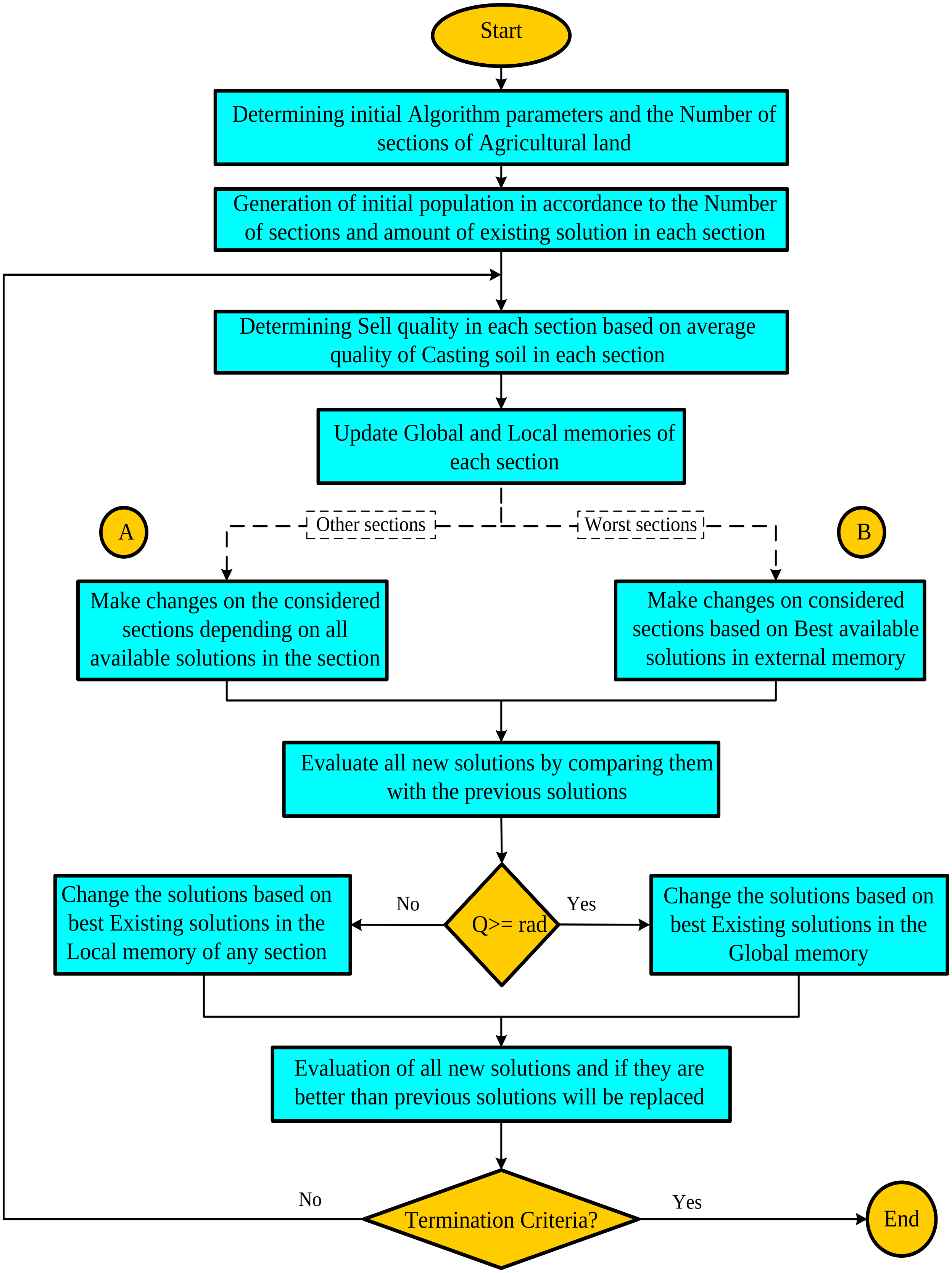

Start
Determining initial Algorithm parameters and the Number of sections of Agricultural land
Generation of initial population in accordance to the Number of sections and amount of existing solution in each section
Determining Sell quality in each section based on average quality of Casting soil in each section
Update Global and Local memories of each section
A
Other sections
Worst sections
B
Make changes on the considered sections depending on all available solutions in the section
Make changes on considered sections based on Best available solutions in external memory
Evaluate all new solutions by comparing them with the previous solutions
Change the solutions based on best Existing solutions in the Local memory of any section
No
Q>= rad
Yes
Change the solutions based on best Existing solutions in the Global memory
Evaluation of all new solutions and if they are better than previous solutions will be replaced
No
Termination Criteria?
Yes
End

Fig. 5. Flowchart of Farmland Fertility Algorithm.

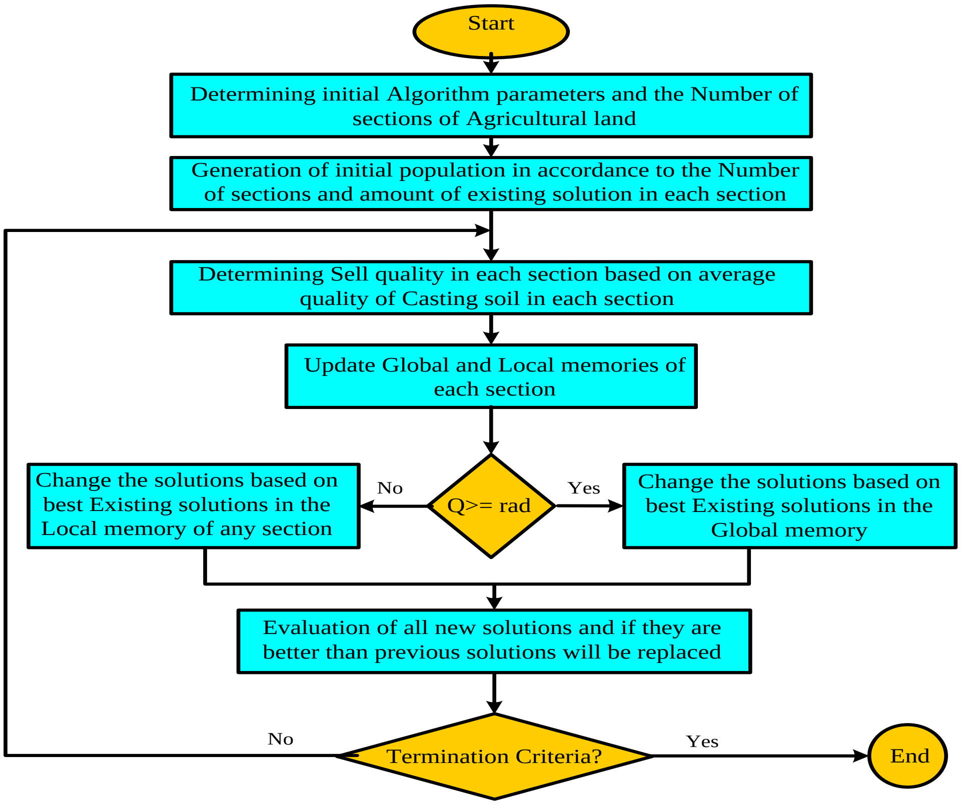

Start
Determining initial Algorithm parameters and the Number of sections of Agricultural land
Generation of initial population in accordance to the Number of sections and amount of existing solution in each section
Determining Sell quality in each section based on average quality of Casting soil in each section
Update Global and Local memories of each section
Change the solutions based on best Existing solutions in the Local memory of any section
No
Q>= rad
Yes
Change the solutions based on best Existing solutions in the Global memory
Evaluation of all new solutions and if they are better than previous solutions will be replaced
No
Termination Criteria?
Yes
End

Fig. 6. Flowchart of Modified Farmland Fertility Algorithm.

V. Results and Discussion

To evaluate the advantages of the proposed optimization algorithm, a real case of study has been presented in order to design a hybrid renewable energy system. The case study has been selected in Ataka, Suez, Egypt. The annual load curve of the region under study is given in Fig. 7. The data of wind speeds over the year and horizontal solar radiation from 7 A.M to 16 P.M are obtained from the Egyptian Metrological Authority for Ataka site. A sample of the input data characterizing the meteorological data of the site including the intensity of solar radiation and wind speed are given in Fig. 8.

The results were accomplished using MATLAB simulation program. The parameters of the optimization algorithms FFA and MFFA are the same; the maximum number of iterations was set to 100 iterations and the maximum number of search agents is 30. In this study, the size of the proposed hybrid system is defined as the number of PV modules, the number of wind turbines, the rated power of the electrolyzer, the mass of hydrogen tank, and the rated power of the fuel cell.

Fig. 9 shows the convergence curves for the FFA and MFFA optimization algorithms. As seen from Fig. 9, both optimization techniques can reach the optimum solution of the objective function. However, the figure shows that the MFFA is faster than the FFA optimization technique. The MFFA can reach to the optimum solution in 8 iteration while the FFA algorithm reach after 11 iteration. Moreover, Table II shows the detailed results of the optimization process for the two optimization algorithms. The essential term for comparison is not only the number of iterations but also the time elapsed for implementation. The implementation time with the proposed MFFA is 3664.6 second for the 100 iteration but the time required for execution the 100 iterations with FFA is 6792.7 second. The main reason for this reduction of the implementation time is due to the conventional FFA requires 2 steps for evaluating the new solution for each iteration while the proposed MFFA does it for one step.

According to FFA optimization algorithms, to ensure the minimum COE of 0.3570 at the proposed location 70 PV units, 100 wind turbines, the rated power of the Electrolyzer of 500kW, the mass of the hydrogen tank of 70kg, and FC with the rated power of 133.4 kW is estimated.

From the results, it is inferred that the both FFA and MFFA predicts minimum COE of 0.3570 \$/kWh, which results in a net present value of 8.6204 million \$ and ensure the value of LPSP of 1.12e-^{19} which agree with the predefined value ($<\beta_L$= 0.05).

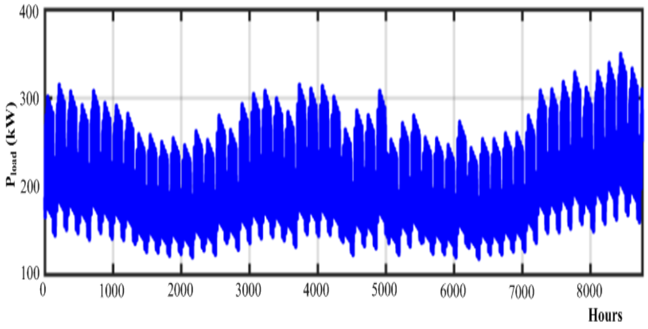
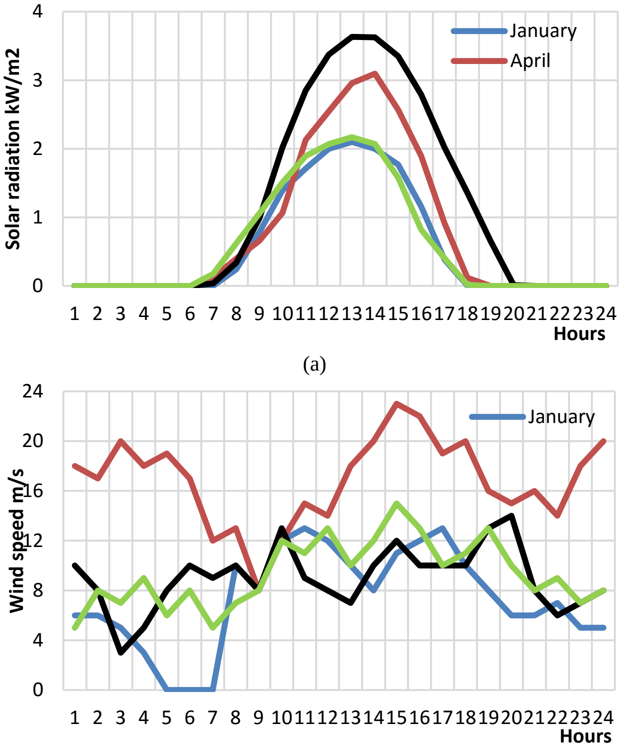


Fig. 7. Load Demand Over a Year.



(a)

(b)

Fig. 8. Annual Variation of Input Parameters; (a) Solar Irradiation, (b) Wind Speed.

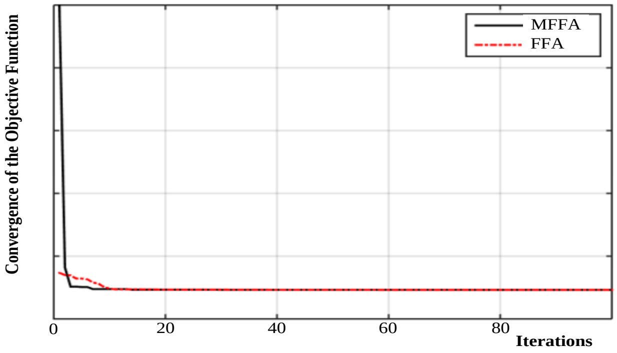


Fig. 9. Convergence Curves for the FFA and MFFA Optimization Techniques.

Fig. 10 shows the hourly variation of the generated power for the components of the proposed hybrid system at the optimum case for MFFA. The presented results from the figures are; the difference between the renewable generation and the load (P_{diff}), the power consumed by the electrolyzer (P_{ELEC}) to convert the excess electrical energy into chemical energy in the form of hydrogen, the mass of the hydrogen stored in the tanks.

(M_{tank}), the power generated by the fuel cell (P_{FC}) in the case of low wind speed and insufficient solar radiation, the electric power transformed from DC form into AC power through the DC/AC converter (P_{INV}), and finally the energy exchange with the external grid, which is represented in the form of surplus and deficit power during the operation period (P_{sur} &P_{def}).

TABLE II. RESULTS OF THE SYSTEM DESIGNED BASED ON FFA AND MFFA

		FFA 2018	MFFA 2019
Best Objective Function		0.3570	0.3570
Best solution	***PV (units)***	70	70
	Wind (units)	100	100
	Electrolyzer (kW)	500	500
	Hydrogen tank (kg)	70	70
	Fuel cell (kW)	133.4	133.4
	DC/AC converter (kW)	500	500
No of iterations for the optimal solution		11	8
COE		0.3570	0.3570
LPSP		1.12e-19	1.12e-19
Time of implantation (second) for 100 iterations		6792.7	3664.6

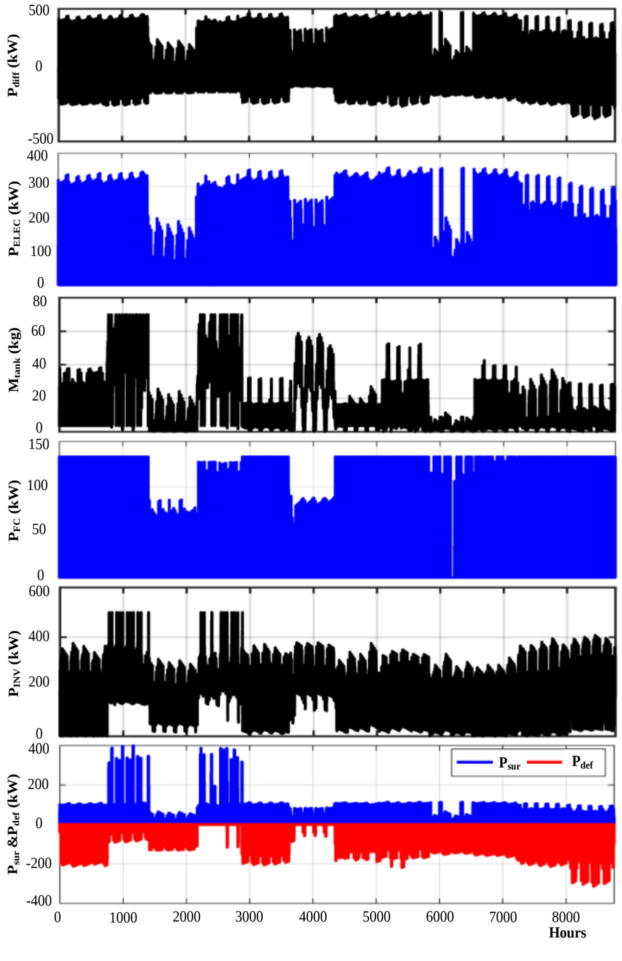


Fig. 10. MFFA Algorithm Results for the Optimal Solution.

Due to design constraints, it is very difficult to achieve the optimization requirements while keeping zero energy exchange with the grid. From the figure, it is clearly seen that almost time electric power flows to and from the external grid to cover the load demand. During the hours of high generation from the renewable sources, the excess energy is used by the electrolyzer and in turn, the mass of hydrogen in the tank increased. For a better understanding of the energy management strategy behind the optimization algorithms, the simulation results are concentrated for one day of operation of the hybrid system at the optimum conditions of operation as shown in Fig. 11.

Fig. 11 shows the simulation results for a certain day with respect to the optimal sizing from MFFA algorithm. From the figure, it is obvious that during late nighttime and the early hours, when the wind speed reaches it maximum values and the load demand is minimum, the generated power from the renewable sources exceeds the demand for electric energy. As a result of that, the excess power is drawn by the electrolyzer, which converts it into hydrogen. At the late hours of night of the day it is very clear from Fig. 11 that M_{tank} increases. This scenario is repeated again during day hours when the solar radiation is maximum. During daytime when the demand for electric energy increased but the renewable generation is not enough to satisfy the load, the hydrogen stored is consumed by the fuel cell. It is obvious from Fig. 11 that M_{tank} reached a maximum value then started to decrease and at that time the power generated by the fuel cell (P_{FC}) started to increase.

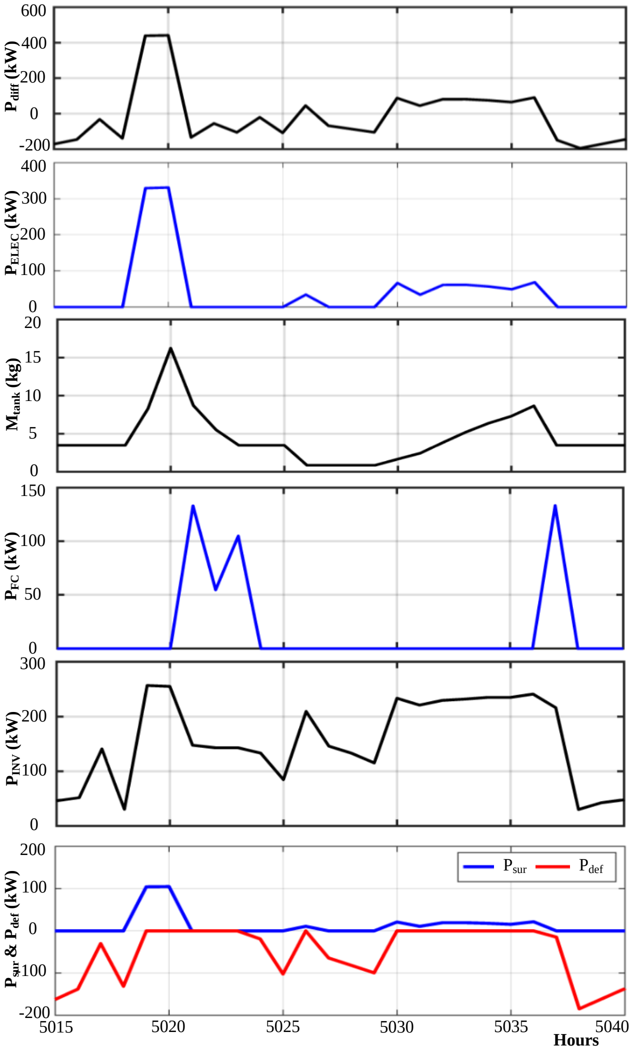


Fig. 11. Simulation Results for One Day of Operation with MFFA.

VI. Conclusion

The optimal sizing and operation of a grid-connected hybrid energy system are performed. For this work, solar photovoltaic-wind turbine-fuel cell are designed and economic feasibility model is introduced through a hybrid grid-connected system. Modified Farmland Fertility Algorithm has been proposed for solving the optimization problem. The PV-wind-electrolyzer-Hydrogen tank-Fuel cell system has been designed to meet the load demand with the minimum COE while ensuring a high power supply reliability, low fluctuations in the energy exchange with the external grid, and complementary use of renewable energy sources. The optimization results show that both FFA and MFFA algorithms reached the best value of the objective function of 0.357, which represents the minimum value of the COE of 0.357 $/kWh. However, MFFA reached the best objective function faster than FFA algorithm. MFFA algorithm reached the minimum COE within 8 iterations, which takes an implementation time of 3664.6 s, while FFA reached the optimal solution in 11 iterations, which takes a time of 6792.7 s. This study will be useful for the decision makers in Egypt as a convenient solution to increase the penetration level of irregular renewable energy sources and ensuring fulltime energy supply. The next research area is to explore the role that high penetration level of large-scale PV power plants will play in short and long-term frequency stability. In addition, in hybrid PV/wind power systems the dynamic nature of these renewable sources has to be fully utilized.

References

[1] Goedeckeb, M.; Therdthianwong, S.; Gheewala, SH. Life cycle cost analysis of alternative vehicles and fuels in Thailand. Energy Policy 2007, 35(6), 3236–3246.

[2] Kusakana, K.; Vermaak, H.J. Hybrid renewable power systems for mobile telephony base station in developing countries. Elsevier Renewable Energy 2013, 51, 419 – 425.

[3] IEA. Renewable energies for remote areas and islands; April, 2012.

[4] Bernal-Agustín, JL; Dufo-López, R. Simulation and optimization of stand-alone hybrid renewable energy systems. Renew Sustain Energy Rev 2009, 13, 2111-8.

[5] Elhadidy, MA; Shaahid, SM. Parametric study of hybrid (wind + solar + diesel) power generating systems. Renew Energy 2000, 21, 129-139.

[6] Bagul, AD; Salameh, ZM; Borowy, B. Sizing of a stand-alone hybrid wind photovoltaic system using a three-event probability density approximation. Sol Energy 1996, 56, 323-335.

[7] Ghosh, GC; Emonts, B; Stolen, D. Comparison of hydrogen storage with diesel generator system in a PV–WEC hybrid system. Sol Energy 2003, 75, 187–198.

[8] Bak, T.; Nowotny, J.; Rekas, M.; Sorrell, CC. Photo-electrochemical hydrogen generation from water using solar energy: materials-related aspects. Int J Hydrogen Energy 2002, 27, 991–1022.

[9] Kashefi, Kaviani A; Baghaee, HR; Riahy, GH. Design and optimal sizing of a photovoltaic/wind-generator system using particle swarm optimization. In Proceedings of the 22nd power system conference (PSC), Tehran, Iran, December 19–21, 2007.

[10] Khan, MJ; Iqbal, MT. Dynamic modeling and simulation of a small wind-fuel cell hybrid energy system. Renewable Energy 2005, 30, 421–439.

[11] Khan, MJ; Iqbal, MT. Pre-feasibility study of stand-alone hybrid energy systems for applications in Newfoundland. Renewable Energy 2005, 30, 835–854.

[12] Mills, A; Al-Hallaj, S. Simulation of hydrogen-based hybrid systems using hybrid2. Int J Hydrogen Energy 2004, 29, 991–999.

[13] Bernal-Agust´ın, J. L.; Dufo-L´opez, R.; Rivas-Ascaso, D. M. Design of isolated hybrid systems minimizing costs and pollutant emissions. Renewable Energy 2006, 31(14), 2227– 2244.

[14] Dufo-L´opez, R.; Bernal-Agust´ın, J.; Yusta-Loyo J. Multiobjective optimization minimizing cost and life cycle emissions of stand-alone PV-wind-diesel systems with batteries storage. Applied Energy 2011, 88(11), 4033–4041.

[15] Yang, H.; Zhou, W.; Lu, L.; Fang, Z. Optimal sizing method for stand-alone hybrid solar–wind system with LPSP technology by using genetic algorithm. Elsevier Journal of Solar Energy 2008, 82, 354-367.

[16] Bilal, B. O.; Sambou, V.; Ndiaye, P. A.; Kebe, C.; Ndongo, M. M. Optimal design of a hybrid solar–wind-battery system using the minimization of the annualized cost system and the minimization of the loss of power supply probability (LPSP). Elsevier Journal of Renewable Energy 2010, 35, 2388-2390.

[17] Tafreshi, M. S.; Zamani, H. A.; Ezzati, S. M.; Baghdadi, M. Optimal unit sizing of Distributed Energy Resources in Micro Grid using genetic algorithm. In 18th Iranian Conference on Electrical Engineering (ICEE), Isfahan, Iran, 2010.

[18] Pirhaghshenasvali, M.; Asaei, B. Optimal modeling and sizing of a practical hybrid wind/PV/diesel generation system. In 5th Power Electronics, Drive Systems and Technologies Conference (PEDSTC), Tehran, Iran, 2014.

[19] Bashir, M.; Sadeh, J. Optimal sizing of hybrid wind/photovoltaic/battery considering the uncertainty of wind and photovoltaic power using Monte Carlo. In Environment and Electrical Engineering (EEEIC), 11th International Conference on, Venice, 2012.

[20] Navaerfard; Tafreshi, S.; Barzegari, M.; Shahrood, A. Optimal sizing of distributed energy resources in microgrid considering wind energy uncertainty with respect to reliability. In IEEE International Energy Conference and Exhibition (EnergyCon), Manama, 2010.

[21] Ekren, O.; Ekren, B. Y. Size optimization of a PV/wind hybrid energy conversion system with battery storage using simulated annealing. Journal of Applied Energy 2010, 87(2), 592-598.

[22] Ekren, O.; Ekren, B. Y. Size optimization of a PV/wind hybrid energy conversion system with battery storage using response surface methodology. Journal of Applied Energy 2008, 85(11), 1086 - 1101.

[23] Arabali, M. Ghofrani, M. Etezadi-Amoli and M. S. Fadali, "Stochastic Performance Assessment and Sizing for a Hybrid Power System of Solar/Wind/Energy Storage," IEEE Transactions on Sustainable Energy 2014, 5(2), 363-371.

[24] Mohamed, Mohamed; Eltamaly, Ali; & Alolah, Abdulrahman; Hatata, Y. A. A novel framework-based cuckoo search algorithm for sizing and optimization of grid-independent hybrid renewable energy systems. International Journal of Green Energy 2018, 1-15. 10.1080/15435075.2018.1533837.

[25] Makbul A.M. Ramli , H.R.E.H. Bouchekara , Abdulsalam S. Alghamdi, “Optimal sizing of PV/wind/diesel hybrid microgrid system using multi-objective self-adaptive differential evolution algorithm,” Renewable Energy 121 (2018) 400-411.

[26] Raviprabakaran, Dr.Vijay; Ramachandradurai, Sowmya. Optimal Sitting of PV-Wind-Energy Storage System Integrated Micro Grid Using Artificial Bee Colony Optimization Technique. International Journal of Innovative Research in Computer and Communication Engineering 2017, 5, 9640-9652.

[27] Ngan, M.; Tan, C. Assessment of economic viability for PV/wind/diesel hybrid energy system in southern Peninsular Malaysia. Renewable and Sustainable Energy Reviews 2012, 16, 634–647.

[28] Haidar, A.; John, P.; Shawal, M. Optimal configuration assessment of renewable energy in Malaysia. Renewable Energy 2011, 36, 881–888.

[29] Sultan, Hamdy M.; Diab, Ahmed A. Zaki; Kuznetsov, Oleg N.; Zubkova, Irina S. Design and evaluation of PV-wind hybrid system with hydroelectric pumped storage on the National Power System of Egypt. Global Energy Interconnection 2018, 1(3), 301-311,

[30] Koutroulis, E; Kolokotsa, D; Potirakis, A; Kalaitzakis, K. Methodology for optimal sizing of stand-alone photovoltaic/wind-generator systems using genetic algorithms. Sol Energy 2006, 80, 1072–1088.

[31] Garcia, RS; Weisser, D. A wind–diesel system with hydrogen storage: joint optimization of design and dispatch. Renewable Energy 2006, 31, 2296–2320.

[32] El-Sharkh, MY; Tanrioven, M; Rahman, A; Alam, MS. Cost related sensitivity analysis for optimal operation of a grid-parallel PEM fuel cell power plant. J Power Sources 2006, 161, 1198–1207.

[33] Kashefi, Kaviani; Riahy, G.H.; Kouhsari, SH.M. Optimal design of a reliable hydrogen-based stand-alone wind/PV generating system, considering component outages. Renewable Energy 2009, 34, 2380–2390.

[34] Strunz, K; Brock, EK. Stochastic energy source access management: infrastructure- integrative modular plant for sustainable hydrogen-electric cogeneration. Int J Hydrogen Energy 2006, 31, 1129–1141.

[35] Xu, L; Ruan, X; Mao, C; Zhang, B; Luo, Y. An improved optimal sizing method for wind–solar–battery hybrid power system. IEEE Trans Sustain Energy 2013; 4(3), 774–785

[36] Shayanfar, Human; Gharehchopogh, Farhad Soleimanian. Farmland fertility: A new metaheuristic algorithm for solving continuous optimization problems. Applied Soft Computing 2018, 71, 728-746.

[37] The official website of the Ministry of Electricity and Renewable Energy, New Renewable Energy Authority, Available at http://www.nrea.gov.eg/Technology/ProjectLocation.

Smart Rubric-based Systematic Model for Evaluating and Prioritizing Academic Practices to Enhance the Education Outcomes

Mohammed Al-Shargabi
Department of Information Systems
Najran University, Najran, Kingdom of Saudi Arabia

***Abstract*—Recently, the impact of free-market economy, globalization, and knowledge economy has become a challenging and focal to higher educational institutions, which resulted in radical change. Therefore, it became mandatory for the academic programs to prepare highly qualified graduates to meet the new challenges, through the implementation of well-defined academic standards. For this reason, the National Center for Academic Accreditation & Evaluation (NCAAA) in Kingdom of Saudi Arabia (KSA) defined a set of standards to ensure that quality of education in KSA is equivalent to the highest international standards. NCAAA standards contains of good criterions to guide the universities in evaluating their quality performance for improvement and obtain NCAAA accreditation. However, implementing NCAAA standards without supportive systems has been found to be a very complex task due to the existence of a large number of standard criterions, evaluation process occurs according to personal opinions, the lack of quality evaluation expertise, and manual calculation. This, in turn, leads to inaccurate evaluation, develops inaccurate improvement plans, and difficulty in obtaining NCAAA accreditation. Therefore, this paper introduces a systematic model that contain smart-rubrics that has been designed based on NCAAA quality performance evaluation elements supported with algorithms and mathematical models to reduce personal opinions, provide an accurate auto-evaluation, and auto-prioritization action plans for NCAAA standards. The proposed model will support academics and administrative by facilitating their NCAAA quality tasks with ease, an authenticate self-assessment, accurate action plans and simplifying accreditation tasks. Finally, the implementation of the model proved to have very efficient and effective results in supporting KSA education institution in accreditation tasks that will lead to enhance the quality of education and to obtain NCAAA accreditation.**

***Keywords*—Systematic Model; Smart Rubric; NCAAA Good Practices; quality of academic programs and universities; improvement action plan**

I. Introduction

The rapid growth of global the free-market economy, globalization, and knowledge economy has created a global competition in higher educational institutions [1-4]. Thus, educational institutions are participating in meeting the high demands of the market and keeping abreast of current technological developments. Therefore, those educational institutions are required to prepare highly qualified graduates who are competent with the needs of the global free-market economy, globalization, and knowledge economy. NCAAA has defined academic standards in 2009 [5-6] and redefined these standards in December 2018 [7-9] to guarantee that Saudis universities and academic programs are qualified for the current challenges. NCAAA standards (both versions) contain good criterions/ practices to guide the academic programs and universities in assessing the competence and usefulness of the educational process, and to use this information to make decisions about how essential activities are enhanced, organized, and funded. Thus, implementation of NCAAA standards by KSA universities and academic programs will ensure good academic performance to meet the current education challenges. NCAAA standards cover different aspects of activities carried out by any academic entity. NCAAA standards are broken down into sub- standards dealing with requirements within each of the major areas. Each of the sub-standards consists of several good criterions/ practices. NCAAA 2009 standards practices at institutions level are more than four-hundred fifty, and at the program level is more than two hundred eighty. While, NCAAA 2018 standards practices at institutions level is one hundred fifty-six (156), at the program level is ninety-six (96), and at postgraduate program level is one-hundred fourteen (114). Currently, NCAAA accredited Saudis universities and academic programs using both (2009, 2018) version (for a specified period of time) leaving the option for universities and academic programs to use any version. NCAAA accepts the accreditation only when the institution has obtained a specific performance level in each standard. High performance level in the standards can only be achieved by accurate, valid and reliable evaluation of performance level and the creation of correct improvement action plans. Therefore, implementation, evaluation, prioritization and construction improvement action plans to achieve a good performance level in NCAAA standards criterion/ practices became a hard task without smart systematic aids and accurate evaluation tools. Hence, this paper will develop two evaluation rubrics (which is an evaluation tool that indicates success criteria to assess different kinds of academic works [10]) to evaluate both versions of NCAAA standards and criterion accurately. Moreover, this paper proposes mathematical equation that will be integrated in algorithm model to develop a smart rubric-based systematic model to auto-evaluate and auto-prioritize evaluate both versions of NCAAA criterions/ practices to support academics and administrative in their planning, self-review, and quality improvement strategies.

This paper is organized as follows: Section II gives an overview of current system for evaluating NCAAA standards, Section III describes the designing of smart rubric-based systematic model for evaluating and prioritizing academic practices to enhance the education outcomes, Section IV describes the practical implementation of the model, and Section V ends with conclusive remarks.

II. Current System for Evaluating NCAAA Standards

Currently, Saudi universities and academic programs use NCAAA standards as guidance for developing, managing, evaluating, and enhancing education programs. NCAAA has defined academic standards in 2009 and improved these standards in December 2018 leaving the option (for a specified period of time) for universities and academic programs to apply for NCAAA accreditation using either version to facilitate the accreditation process for the institutions who build their quality systems based on 2009 standards.

NCAAA 2009 standards consist of 11 broad standards that apply to both institutions and programs though there are differences in how they are applied for these different kinds of evaluations. NCAAA 2009 standards 11 standards are:

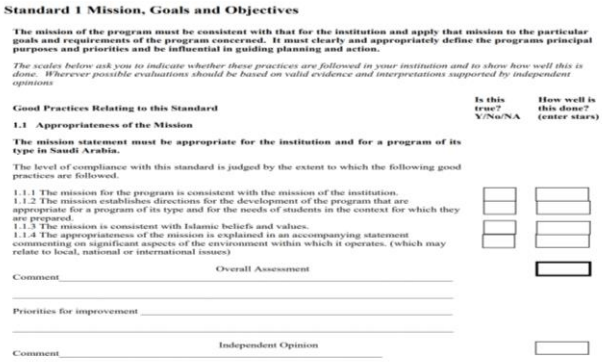
(1) Mission and Objectives, (2) Governance and Administration (3) Management of Quality Assurance and Improvement, (4) Learning and Teaching, (5) Student Administration and Support Services, (6) Learning Resources, (7) Facilities and Equipment, (8) Financial Planning and Management, (9) Faculty and Staff Employment Processes, (10) Research and (11) Relationship with the Community.

NCAAA has prepared 2009 Self-Evaluation Scales (SES) document to help Saudi universities and academic programs to evaluate the NCAAA 2009 standards for quality level. SES support higher education institutions in enhancing their ability to meet the standards of quality assurance and to be used in NCAAA academic accreditation. SES is used by institutions in self-initial quality assessment, continues improvement plans, and prepares a self-study report to obtain NCAAA accreditation. Currently, SES standards evaluation is conducted manually by collecting the points of evaluation for all the related criteria according to their quality performance in elements of evaluation.

NCAAA has prepared two documents for NCAAA 2009 SES (sample is shown in Fig. 1) which is SES for higher education institution [11], and SES for higher education programs [12] (in MS-Word, and PDF format) to evaluate quality performance of NCAAA practices.

NCAAA 2009 SES has three elements of evaluation, which are: the extent and consistency with which processes are followed, the quality of the service or activity as assessed through systematic evaluations; and the effectiveness of what is done in achieving intended outcomes.

NCAAA 2009 SES evaluates the standards by categorizing the applicable practice quality performance into three performance level which are low, good, and high performance using zero to five stars evaluation system as shown in Fig. 2.



Standard 1 Mission, Goals and Objectives

The mission of the program must be consistent with that for the institution and apply that mission to the particular goals and requirements of the program concerned. It must clearly and appropriately define the programs principal purposes and priorities and be influential in guiding planning and action.

The scales below ask you to indicate whether these practices are followed in your institution and to show how well this is done. Wherever possible evaluations should be based on valid evidence and interpretations supported by independent opinions

Good Practices Relating to this Standard	**Is this true? Y/No/NA**	**How well is this done? (enter stars)**
1.1 Appropriateness of the Mission		
The mission statement must be appropriate for the institution and for a program of its type in Saudi Arabia.		
The level of compliance with this standard is judged by the extent to which the following good practices are followed.		
1.1.1 The mission for the program is consistent with the mission of the institution.	□	□
1.1.2 The mission establishes directions for the development of the program that are appropriate for a program of its type and for the needs of students in the context for which they are prepared.	□	□
1.1.3 The mission is consistent with Islamic beliefs and values.	□	□
1.1.4 The appropriateness of the mission is explained in an accompanying statement commenting on significant aspects of the environment within which it operates. (which may relate to local, national or international issues)	□	□
Overall Assessment		□

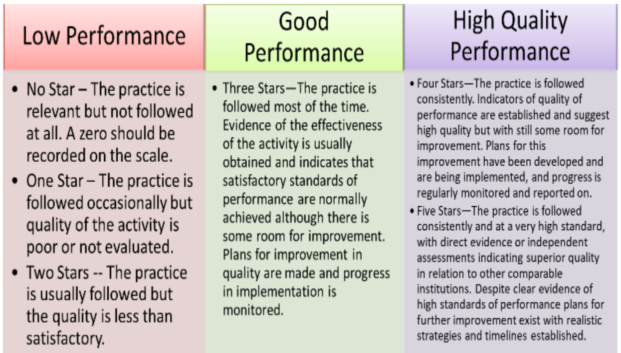
Comment______________________________

Priorities for improvement ________________________

Independent Opinion □

Comment__________________________

Fig. 1. Simple of NCAAA 2009 SES for Higher Education Programs Templates.



Low Performance	Good Performance	High Quality Performance
• No Star – The practice is relevant but not followed at all. A zero should be recorded on the scale. • One Star – The practice is followed occasionally but quality of the activity is poor or not evaluated. • Two Stars -- The practice is usually followed but the quality is less than satisfactory.	• Three Stars—The practice is followed most of the time. Evidence of the effectiveness of the activity is usually obtained and indicates that satisfactory standards of performance are normally achieved although there is some room for improvement. Plans for improvement in quality are made and progress in implementation is monitored.	• Four Stars—The practice is followed consistently. Indicators of quality of performance are established and suggest high quality but with still some room for improvement. Plans for this improvement have been developed and are being implemented, and progress is regularly monitored and reported on. • Five Stars—The practice is followed consistently and at a very high standard, with direct evidence or independent assessments indicating superior quality in relation to other comparable institutions. Despite clear evidence of high standards of performance plans for further improvement exist with realistic strategies and timelines established.

Fig. 2. Simple of NCAAA 2009 SES for Higher Education Programs Templates.

Higher educational institutions and programs use 2009 SES templates to calculate manually the quality performance level of each practice, using zero to five stars evaluation system based on the evaluation of the practice. Then, higher education institution and programs manually calculate the evaluation stars of each sub-standard by taking the average for all the practices in that sub-standard. Finally, higher education institution and programs calculate manually the evaluation stars of each standard by taking the average of for all sub-standards in that standards. Based on the evaluation of each standard, higher education institution and programs prepare an improvement plan to enhance the quality of the university/ program.

However, implementing the above-given evaluation system to evaluate and enhance NCAAA standards and practices is not an easy task due to the large number of practices, personal opinions-based evaluation process, lack of quality evaluation expertise, and the difficulty of manual calculation. Moreover, the absence of indicators for NCAAA practices priorities and importance leads to inaccurate improvement plans, which leaves the institution and/or the programs without an actual continuous improvement process.

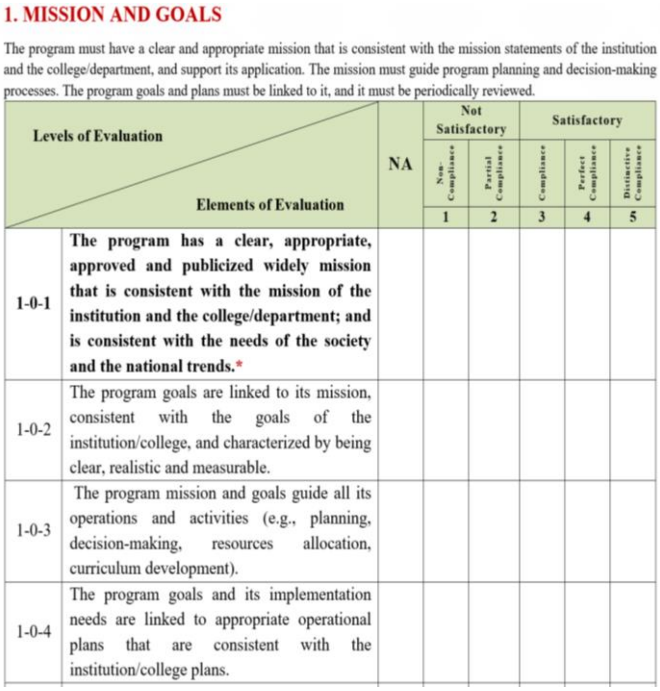
Therefore, NCAAA has redefined NCAAA 2009 standards in December 2018 to facilitate its accreditation tasks and overcome some of NCAAA 2009 standards evaluation system with giving the option (for a specified period of time) for universities and academic programs to apply for accreditation using 2099 NCAAA standards.

By December 2018, NCAAA introduced an improved version of NCAAA standards to be eight standards at the institutions level, six standards at the program level, and seven standards for postgraduate programs. NCAAA 2018 institutional standards, which are:

(1) Mission, Goals and Strategic Planning, (2) Governance, Leadership and Management, (3) Teaching and Learning, (4) Students, (5) Faculty and Staff, (6) Institutional Resources, (7) Scientific Research and Innovation, and (8) Community Partnership while, NCAAA 2018 program standards [9] are: (1) Mission and goals, (2 Program management and quality assurance, (3) Teaching and Learning, (4) Students, (5) Faculty members, and (6) Learning resources, facilities, and equipment. Newly, NCAAA 2018 proposed a specific standard for postgraduate programs [10] which are: (1) Mission and goals, (2) Program management and quality assurance, (3) Teaching and Learning, (4) Students, (5) Faculty members, (6) Learning resources, facilities, and equipment, and (7) Research and Projects.

Also, NCAAA has prepared two documents for NCAAA 2018 SES (sample is shown in Fig. 3) which is SES for higher education institution [13], and SES for higher education programs [14] to evaluate quality performance of NCAAA improved criterions.

NCAAA 2018 SES has five elements of evaluation which are: extent of availability of elements and components of the criterion, quality level of application for each element, regularity of application and assessment, and availability of evidence, continuous improvement and level of results in the light of indicators and benchmarks, excellence and creativity in practices of the elements of the criterion. NCAAA 2018 SES improved a guidance rubric (not complete rubric for all NCAAA 2018 criterion) as shown in Fig. 4.



1. MISSION AND GOALS

The program must have a clear and appropriate mission that is consistent with the mission statements of the institution and the college/department, and support its application. The mission must guide program planning and decision-making processes. The program goals and plans must be linked to it, and it must be periodically reviewed.

Levels of Evaluation / Elements of Evaluation		NA	Not Satisfactory		Satisfactory		
			Non-Compliance	Partial Compliance	Compliance	Perfect Compliance	Distinctive Compliance
			1	2	3	4	5
1-0-1	**The program has a clear, appropriate, approved and publicized widely mission that is consistent with the mission of the institution and the college/department; and is consistent with the needs of the society and the national trends.***						
1-0-2	The program goals are linked to its mission, consistent with the goals of the institution/college, and characterized by being clear, realistic and measurable.						
1-0-3	The program mission and goals guide all its operations and activities (e.g., planning, decision-making, resources allocation, curriculum development).						
1-0-4	The program goals and its implementation needs are linked to appropriate operational plans that are consistent with the institution/college plans.						

Fig. 3. Simple of NCAAA 2018 SES for Higher Education Programs Templates.

Levels of Evaluation / Elements of Evaluation	NA	Unsatisfactory Performance		Satisfactory Performance		
		Non-Compliance	Partial Compliance	Compliance	Perfect Compliance	Distinctive Compliance
		1	2	3	4	5
Extent of availability of elements and components of the criterion		• There are no available elements of the criterion • Or there are few available elements	• Most of the elements of the criterion are available	• All of the elements of the criterion are available	• All of the elements of the criterion are available	• All of the elements of the criterion are available
Quality level of application for each element		• The elements of the criterion are not applied at all, (or) are applied at a very low level	• The elements of the criterion are applied at low level	• The elements of the criterion are applied at good level	• The elements of the criterion are applied at perfect level	• The elements of the criterion are applied at distinct level
Regularity of application and assessment, and availability of evidence		• Rarely applied	• Applied irregularly, • (or) there is no assessment, or it is there but is irregular, • (or) there is insufficient evidence	• Applied regularly, • There is a regular and effective assessment, • Sufficient evidence is available	• Applied regularly, • There is a regular and effective assessment, • Sufficient and varied evidence is available	• Applied on a regular basis, • There is a regular, effective, and excellent assessment, and • Various, comprehensive, and cumulative evidence is available,

Fig. 4. Simple of NCAAA 2018 SES Guidance.

NCAAA 2018 SES evaluates the standards by categorizing the applicable criterion quality performance into two performance levels which are unsatisfactory performance, and satisfactory performance using a five-point evaluation scale (1 to 5) as shown in Fig. 5.

SES 2018 templates will be filled by evaluating each criterion performance level and giving a number manually, using a five-points scale. Then, each sub-standard performance level will be calculated manually by the average of its criterion' points (if the standard has sub-standards). Finally, each standard performance level will be calculated manually as the average of its sub-standards' points (if the standard has sub-standards) and as the average of its criterion' points (if the standard has no sub-standards). According to the evaluation of each standard, an improvement plan will be developed to enhance the standards.

Level	Description of Performance
(1) **Non-Compliance**	The program does not have teaching and learning strategies, and assessment methods to develop the students' ability to conduct scientific research, and to acquire higher thinking and self-learning skills, or they exist but are inappropriate or incompatible with the nature and level of the program, or that they are not fully applied or are applied rarely or at a very low level.
(2) **Partial Compliance**	The program has limited teaching and learning strategies and assessment methods to develop the students' ability to conduct scientific research, and to acquire higher thinking and self-learning skills, or only some of them are compatible with the nature and level of the program, or they are poorly or irregularly applied, or they are not subject to assessment or some of them are irregularly assessed, and there are limited procedures for their development.
(3) **Compliance**	The program has diverse teaching and learning strategies and assessment methods, compatible with its nature and level, all of which are applied at a good level on a regular basis, for enhancing the ability to conduct scientific research and ensuring students' acquisition of higher thinking and self-learning skills. There is sufficient evidence. Most of them are subject to periodic evaluation and development.
(4) **Perfect Compliance**	The program has diverse and developed teaching and learning strategies and assessment methods, all of which are of a high quality compatible with its nature and level, all of which are applied at a high level on a regular basis, enhancing the ability to conduct scientific research and ensuring students' acquisition of higher thinking and self-learning skills. There is ample and varied evidence. All are subject to periodic evaluation and development with the existence of high results for improvement.
(5) **Distinctive Compliance**	Any distinction and creativity in the practices of the elements of the criterion

Fig. 5. NCAAA 2018 SES Performance Level Description.

The improved NCAAA 2018 standards SES have reduced the efforts of NCAAA accreditation tasks due to the smaller number of criterions compared to NCAAA 2009 standards practices. In addition, the guidance rubric will make the standard evaluation more accurate compared to NCAAA 2009 standards evaluation system. However, the proposed guidance rubric is not a complete rubric for all NCAAA 2018 criterion which will make the evaluation process still based on personal opinions. Moreover, implementing and evaluating NCAAA 2018 standards is still not an easy task due to the number of criterions, lack of complete quality evaluation guidance, the difficulty of manual calculation, and the absence of indicators for criterion priorities for improvement. Therefore, it will be difficult to develop accurate improvement plans.

Thus, there is a need for systematic model for facilitating NCAAA tasks to have more accurate results. Deanship of development and quality in King Saud University (KSU) has an electronic system to manage the process of development and quality, and NCAAA accreditation tasks at the university called ITQAN [15]. ITQAN support KSU academics and administrative with many services such as facilitating access to the data, automating large numbers of periodic reports, and disseminating and analyzing questionnaires. However, ITQAN cannot support in auto-evaluation, and auto-prioritizing the performance level of NCAAA criterion/practices.

Researches in King Abdelaziz University (KAU) propose a system [16] that automates Key Performance Indicators (KPIs) management process for higher educational institutions through balanced scorecard measuring tools. However, this proposed system will support NCAAA KPIs calculations without evaluating NCAAA criterion/practices.

NCAAA developed an electronic accreditation system called DAMAN [17] which will facilitate NCAAA accreditation processes through replacing the traditional paper-based accreditation processes to an integrated electronic accreditation process that saves time, effort and resources. However, DHMAN was developed to facilitate NCAAA accreditation processes not supporting educational institution and programs to evaluate their criterion/practices.

Thus, this paper introduces a smart-rubrics systematic model that is designed to support educational institutions and programs to evaluate their NCAAA criterion/practices, facilitate their NCAAA quality tasks, provide self-assessment, guide in development of accurate quality implementation action plans and simplifying accreditation tasks.

III. Designing of Smart Rubric-Based Systematic Model for Evaluating and Prioritizing Academic Practices to Enhance the Education Outcomes

The proposed smart rubric-based systematic model for evaluating and prioritizing academic practices consists of two evaluation rubrics that are designed to accurately evaluate both versions of NCAAA standards, and an algorithm model that contains mathematical model to auto-evaluate, auto-prioritize, and auto-calculate the performance level of NCAAA standards.

A. Designing the Rubrics

Two rubrics are designed to assess and evaluate academic criterion/practice, one is according to NCAAA 2009 Standards practice, the other one according to NCAAA 2018 Standards criterion.

1) NCAAA 2009 standards rubrics: To evaluate NCAAA 2009 standards practice accurately , a rubric is designed based on three performance criteria: the extent and consistency with which processes are followed, the quality of the service or activity as assessed through systematic evaluations; and the effectiveness of what is done in achieving intended outcomes which are according to NCAAA 2009 standard practice guideline [12, 13]. Each of those performance criteria has its own descriptor aligned with the performance level in the rubric.

Table I shows the rubric that is designed to illuminate the performance criteria, performance descriptor, performance level, the variable's name is ECF, with its possible values (which will be used in the mathematical equations) to evaluate the practices according to the extent and consistency with which processes are followed.

Table II shows the rubric that is designed to illuminate the performance criteria, performance descriptor, performance level, the variable's name is QSA, with its possible values (which will be used in the mathematical equations) to evaluate the practices according to the quality of the service or activity as assessed through systematic evaluations.

Table III shows the rubric that is designed to illuminate the performance criteria, performance descriptor, performance level, the variable's name is EFF, with its possible values (which will be used in the mathematical equations) to evaluate the practices according to the effectiveness of what is done in achieving intended outcomes.

TABLE. I. Rubric Elements to Evaluate the Practices According to the Extent and Consistency with which Processes are Followed

Practice Number and Description	**The extent and consistency with which Processes are Followed**				
	All the time	***Consistently***	***Most of the Time***	***Usually***	***Occasional***
Practice Number and Description	Practice Followed All the time	Practice Followed Consistently	Practice Followed Most of the Time	Practice Followed Usually	Practice Followed Occasionally
ECF	5	4	3	2	1

TABLE. II. Rubric Elements to Evaluate the Practices According to the Quality of the Service or Activity as Assessed through Systematic Evaluations

Practice Number and Description	**The quality of the service or activity as assessed through systematic evaluations**				
	Superior Quality	***High Quality***	***Satisfactory***	***Less than satisfactory***	***Poor***
Practice Number and Description	Practice Quality is Superior Quality	Practice Quality is High Quality	Practice Quality is Satisfactory	Practice Quality is Less than satisfactory	Practice Quality is Poor
QSA	5	4	3	2	1

TABLE. III. RUBRIC ELEMENTS TO EVALUATE THE PRACTICES ACCORDING TO THE EFFECTIVENESS OF WHAT IS DONE IN ACHIEVING INTENDED OUTCOMES

Practice Number and Description	**The effectiveness of what is done in achieving intended outcomes**				
	Excellent	***Very Good***	***Good***	***Satisfactory***	***Poor***
Practice Number and Description	Practice Effectiveness is Excellent	Practice Effectiveness is Very Good	Practice Effectiveness is Good	Practice Effectiveness is Satisfactory	Practice Effectiveness is Poor
EFF	5	4	3	2	1

2) NCAAA 2018 standards rubrics: The NCAAA 2018 Standards criterion rubric is designed based on five performance criteria: extent of availability of elements and components of the criterion, quality level of application for each element, regularity of application and assessment, and availability of evidence, continuous improvement and level of results in the light of indicators and benchmarks, and excellence and creativity in practices of the elements of the criterion according to NCAAA 2018 Standard criterion guideline [14, 15]. Each of those performance criteria has its own descriptor aligned with the performance level in the rubric. Table IV shows the rubric that is designed to illuminate the performance criteria, performance descriptor, performance level, the variable's name is EV, with its possible values (which will be used in the mathematical equations) to evaluate the criterion according to the extent of availability of elements and components of the criterion.

Table V shows the rubric that is designed to illuminate the performance criteria, performance descriptor, performance level, the variable's name is AQ, with its possible values (which will be used in the mathematical equations) to evaluate the criterion according to the quality level of application for each element.

Table VI shows the rubric that is designed to illuminate the performance criteria, performance descriptor, performance level, the variable's name is RA, with its possible values (which will be used in the mathematical equations) to evaluate the criterion according to the regularity of application and assessment, and availability of evidence.

TABLE. IV. RUBRIC ELEMENTS TO EVALUATE THE CRITERION ACCORDING TO THE EXTENT OF AVAILABILITY OF ELEMENTS AND COMPONENTS OF THE CRITERION

Criterion Number and Description	**Extent of availability of elements and components of the criterion**			
	All Elements Available	***Most of the Elements Available***	***Few Available Elements***	***No Available Elements***
Criterion Number and Description	All of Criterion Elements Available	Most of Criterion Elements Available	Few Available Criterion Elements	Criterion Elements Not Available
EV	4	3	2	1

TABLE. V. RUBRIC ELEMENTS TO EVALUATE THE CRITERION ACCORDING TO THE QUALITY LEVEL OF APPLICATION FOR EACH ELEMENT

Criterion Number and Description	**Quality level of application for each element**				
	Applied at Distinct Level	***Applied at Perfect Level***	***Applied at Good Level***	***Applied at Low Level***	***Not Applied at all Very Low Level***
Criterion Number and Description	Criterion Elements are Applied at Distinct Level	Criterion Elements are Applied at Perfect Level	Criterion Elements are Applied at Good Level	Criterion Elements are Applied at Low Level	Criterion Elements Not Applied at all or Applied at a Very Low Level
AQ	5	4	3	2	1

TABLE. VI. RUBRIC ELEMENTS TO EVALUATE THE CRITERION ACCORDING TO THE REGULARITY OF APPLICATION AND ASSESSMENT AND AVAILABILITY OF EVIDENCE

Criterion Number and Description	**Regularity of application and assessment, and availability of evidence**				
	Applied Regularly / Regular Effective, Excellent Assessment / Comprehensive, Cumulative Evidences	***Applied Regularly / Regular and Effective Assessment / Sufficient and Varied Evidences***	***Applied Regularly / Regular and Effective Assessment / Sufficient Evidences***	***Applied Irregularly / (No Assessment (or) Irregular Assessment) / Insufficient Evidences***	***Rarely Applied***
Criterion Number and Description	Criterion Applied Regularly / Regular Effective, Excellent Assessment / Comprehensive, Cumulative Evidences	Criterion Applied Regularly / Regular and Effective Assessment / Sufficient Evidences	Criterion Applied Regularly / Regular and Effective Assessment / Sufficient Evidences	Criterion Applied Regularly / Regular and Effective Assessment / Sufficient Evidences	Criterion Rarely Applied
RA	5	4	3	2	1

Table VII shows the rubric that is designed to illuminate the performance criteria, performance descriptor, performance level, the variable's name is CI, with its possible values (which will be used in the mathematical equations) to evaluate the criterion according to the continuous improvement and level of results in the light of indicators and benchmarks.

TABLE. VII. RUBRIC ELEMENTS TO EVALUATE THE CRITERION ACCORDING TO THE CONTINUOUS IMPROVEMENT AND LEVEL OF RESULTS IN THE LIGHT OF INDICATORS AND BENCHMARKS

Criterion Number and Description	**Continuous improvement and level of results in the light of indicators and benchmarks**			
	Regular Improvement Procedures and Distinct Results Compared To Other Institutions	***Regular Improvement Procedures and Higher Results Compared to Previous Results.***	***Regular Improvement Procedures and Good Results***	***Limited Improvement Procedures***
Criterion Number and Description	Regular Improvement Procedures Applied on the Criterion with Distinct Results Compared To Other Institutions	Regular Improvement Procedures Applied on the Criterion with Higher Results Compared to Previous Results	Regular Improvement Procedures Applied on the Criterion with Good Results	Limited Improvement Procedures Applied on the Criterion
CI	4	3	2	1

TABLE. VIII. RUBRIC ELEMENTS TO EVALUATE THE CRITERION ACCORDING TO THE EXCELLENCE AND CREATIVITY IN PRACTICES OF THE ELEMENTS OF THE CRITERION

Criterion Number and Description	**Excellence and creativity in practices of the elements of the criterion**
	Creativity in the Practices of the Elements of the Criterion.
Criterion Number and Description	There is a Creativity in the Practices of the Elements of the Criterion
EC	5

Table VIII shows the rubric that is designed to illuminate the performance criteria, performance descriptor, performance level, the variable's name is EC, with its possible values (which will be used in the mathematical equations) to evaluate the criterion according to the excellence and creativity in practices of the elements of the criterion.

B. *Designing the Algorithm Model*

An algorithm model is integrated in the rubric to build a smart rubric-based systematic model for evaluating and prioritizing academic practices to enhance the educational outcomes. Fig. 6 shows the algorithm model flowchart. The algorithm model steps can be summarized in the following points:

- The algorithm model will check which type of standards (institutional or program) the user will use. If its institutional standards, the algorithm model will use the mathematical equations of the institutional standards rubrics. Otherwise, the algorithm model will use the mathematical equations of program standards rubrics.

- In both cases in the previous step, the algorithm model will check which version of standards (2018 or 2009) the user will use. If its 2018 standards, the algorithm model will use NCAAA 2018 improved standards rubrics and according to the type of standards (institutional or program) was selected in the previous step. Otherwise, NCAAA 2009 standards rubrics will use according to the type of standards (institutional or program) which was selected in the previous step.

- The algorithm model will use the smart rubric to evaluate criterion/practice according to the selection in the previous steps,

- The algorithm model will use a mathematical equation that is formulated to calculate the criterion/ practice performance evaluation in NP_PerEv according to the selection in the previous steps.

- If the user selects to use NCAAA 2018 improved standards, the following mathematical equations will be used to calculate the criterion points in CP(x) where x is equal to the criterion number:

$$IF\big((EV < 3)\ OR\ (AQ \leq 2)\big)\ then\ \ CP(x) = 1 \tag{1}$$

$$IF\big((EV = 3)\ OR\ (AQ = 2)\ OR\ (RA = 2)\ OR\ (CI = 1)\big)$$

$$then\ CP(x) = 2 \tag{2}$$

$$\text{IF((EV=4) OR (AQ=3) OR (RA=3) OR (CI=2))}$$

$$\text{then CP(x)=3} \tag{3}$$

$$\text{IF((EV=4) OR (AQ=4) OR (RA=4) OR (CI=3))}$$

$$\text{then CP(x)=4} \tag{4}$$

$$\text{IF((EV=4) OR (AQ=5) OR (RA=5) OR (CI=5) OR (EC=5))}$$

$$\text{then CP(x)=5} \tag{5}$$

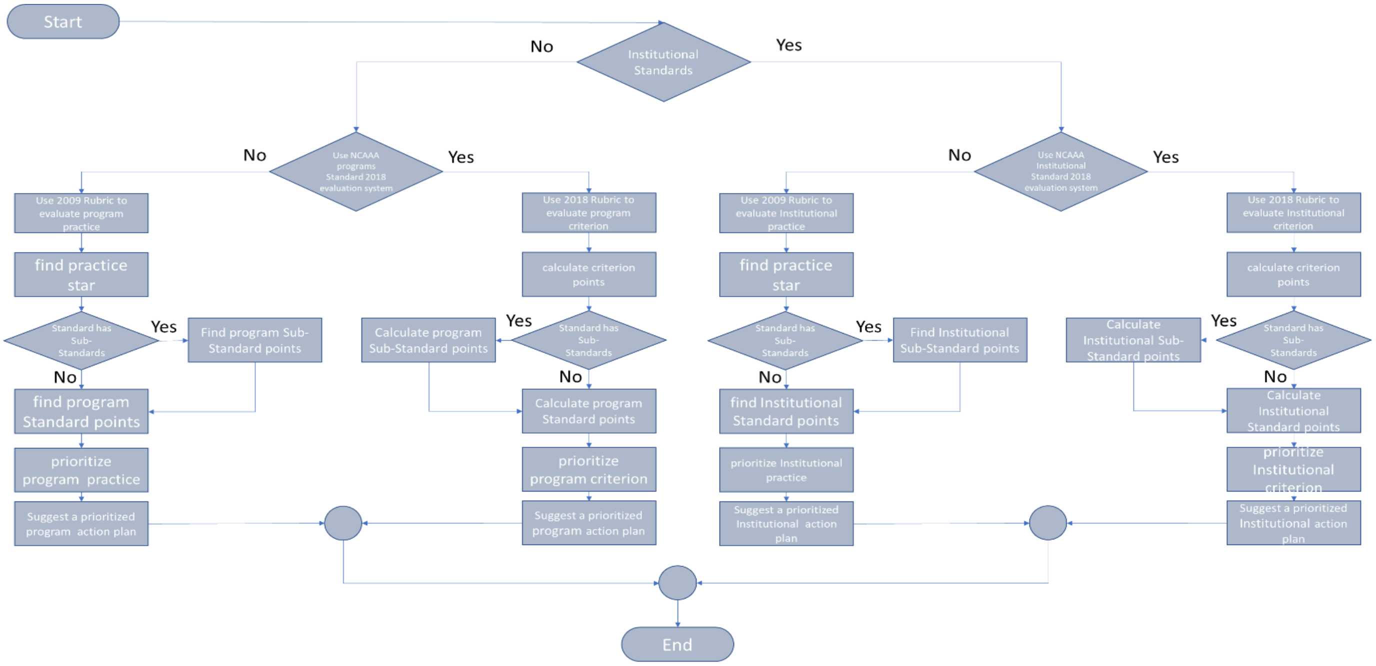

Start
Institutional Standards
No
Yes
Use NCAAA programs Standard 2018 evaluation system
No
Yes
Use 2009 Rubric to evaluate program practice
find practice star
Standard has Sub-Standards
Yes
Find program Sub-Standard points
No
find program Standard points
prioritize program practice
Suggest a prioritized program action plan
Use 2018 Rubric to evaluate program criterion
calculate criterion points
Standard has Sub-Standards
Yes
Calculate program Sub-Standard points
No
Calculate program Standard points
prioritize program criterion
Suggest a prioritized program action plan
Use NCAAA Institutional Standard 2018 evaluation system
No
Yes
Use 2009 Rubric to evaluate Institutional practice
find practice star
Standard has Sub-Standards
Yes
Find Institutional Sub-Standard points
No
find Institutional Standard points
prioritize Institutional practice
Suggest a prioritized Institutional action plan
Use 2018 Rubric to evaluate Institutional criterion
calculate criterion points
Standard has Sub-Standards
Yes
Calculate Institutional Sub-Standard points
No
Calculate Institutional Standard points
prioritize Institutional criterion
Suggest a prioritized Institutional action plan
End

Fig. 6. Smart Rubric-based Systematic Model for Evaluating and Prioritizing Academic Practices Algorithm Model Flowchart.

- If the user selects to use NCAAA 2009 standards, the following mathematical equations will be used to find the practice star in PP(x) where x is equal to the practice number:

$$\text{IF(ECF = 1) then PP(x)=1} \quad (6)$$

$$\text{IF((ECF=2) AND (QSA=2) AND (EFF=1)}$$

$$\text{then PP(x)=2} \quad (7)$$

$$\text{IF((ECF=3) AND (QSA=3) AND (EFF=2))}$$

$$\text{then PP(x)=3} \quad (8)$$

$$\text{IF((ECF} \geqslant \text{4) AND (QSA=4) AND (EFF=3)}$$

$$\text{then PP(x)=4} \quad (9)$$

$$\text{IF((ECF=5) AND (QSA=5) AND (EFF>3)}$$

$$\text{then PP(x)=5} \quad (10)$$

- The algorithm model will check if the evaluated standard has sub-standards, If it does, the algorithm model calculates the sub-standard performance evaluation in SSP(x) according to the selection in the previous steps. Otherwise, the algorithm model moves to calculate the standard performance evaluation.

- The algorithm model will use a mathematical equation that is formulated to calculate the sub-standard points in SSP(x) where x is equal to the sub-standard number and NoP is the number of the criterion in the sub-standard:

$$SSP(x) = \frac{\sum_{n=1}^{NoP} CP(n)}{NoP} \quad (11)$$

- The algorithm model will use a mathematical equation that is formulated to calculate the standard performance evaluation points in SP(x) (where x is equal to the standard number and NoS is the number of the of sub-standard) according to the selection in the previous steps:

IF SSP(x)>0 then

$$SP(x) = \frac{\sum_{n=1}^{NoS} SSP(n)}{NoS}$$

else

$$SP(x) = \frac{\sum_{n=1}^{NoP} SSP(n)}{NoP} \quad (12)$$

- The algorithm model will use the following mathematical equation to auto-prioritize criterion/ practice based on its performance evaluation in PriIM . The algorithm model use the variable ILP to get the importance level of criterion/ practice, if its essential practice, then ILP=1, else ILP=0:

$$IF\big((NP_{PerEv} < 3)\ AND\ (ILP = 1)\big)\ then\ PriIM = 5 \quad (13)$$

$$IF\big((NP_{PerEv} < 3)\ AND\ (ILP = 0)\big)\ then\ PriIM = 4 \quad (14)$$

$$IF\big((NP_{PerEv} = 3)\ AND\ (ILP = 1)\big)\ then\ PriIM = 3 \quad (15)$$

$$IF\big((NP_{PerEv} = 3)\ AND\ (ILP = 0)\big)\ then\ PriIM = 2 \quad (16)$$

$$IF\big((NP_{PerEv} > 3\ )\big)\ then\ PriIM = 1 \quad (17)$$

Where the value 5 means very high priority for improvement, 4 means high priority for improvement, 3 means medium priority for improvement, 2 means normal priority for improvement, and 1 means low priority for improvement.

- The algorithm model will suggest a prioritized action plan according to the selection in the previous steps. The prioritized action plan will contain the criterion/ practice that needs very high priority for improvement, or high priority for improvement according to the selection in the previous steps by implementing the following mathematical equation:

$$IF\ PriIM = 5\ then$$

$$add\ to\ the\ top\ list\ of\ the\ prioritized\ action\ plan$$

$$else\ IF\ PriIM = 4\ then$$

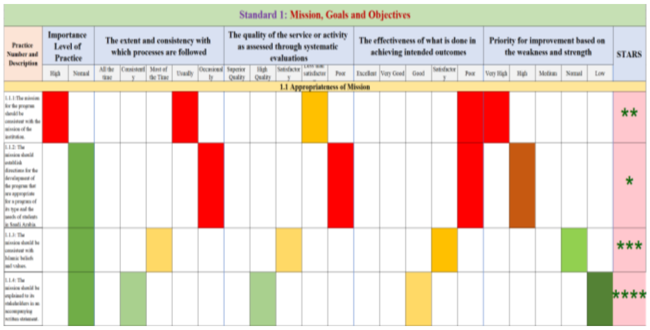
$$add\ to\ the\ bottom\ list\ of\ the\ prioritized\ action\ plan \quad (18)$$

IV. Implementation

The implementation of the smart rubric-based systematic model showed very efficient and effective result in supporting institution and programs in auto-evaluating, auto-prioritizing, and auto-calculating the performance level of NCAAA standards. The proposed model provides a visual and easy selection rubric to support the users to evaluate the criterion/ practice according to the designed rubric in the previous section. When the user selects the performance level of each evaluation element, the smart rubric (as shown in Fig. 7 for NCAAA 2009 standards, and Fig. 8 for NCAAA 2018 standards) can auto-evaluate criterion/ practice, auto-calculate the star/ point, and auto-prioritize and suggest priority for improvement of criterion/ practice.

The smart rubric can support and facilitate academics and administrative workers by suggesting a prioritized accurate action plan according to the criterion/ practice performance evaluation as shown in Fig. 9. The accurate action plan will lead to enhance the university's/ program's quality of education and facilitate the tasks of obtaining NCAAA accreditation.

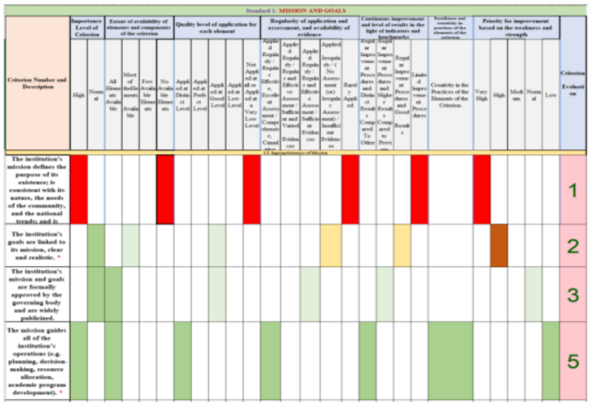
The smart rubric can provide a comparison of the standards performance evaluation (as shown in Fig. 10 for NCAAA 2009 standards, and Fig. 11 for NCAAA 2018 standards) which will support the institutions to easily take decisions for improvement.



Standard 1: Mission, Goals and Objectives

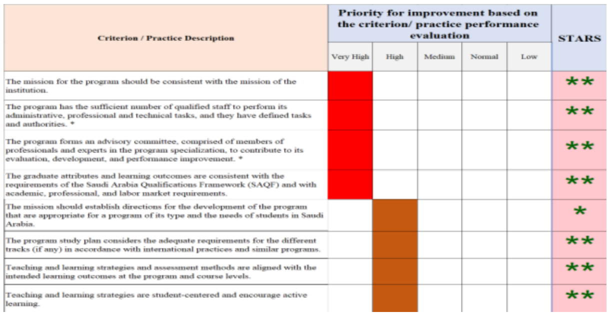
Practice Number and Description	Importance Level of Practice		The extent and consistency with which processes are followed					The quality of the service or activity as assessed through systematic evaluations					The effectiveness of what is done in achieving intended outcomes					Priority for improvement based on the weakness and strength					STARS
	High	Normal	All the time	Consistently	Most of the Time	Usually	Occasionally	Superior Quality	High Quality	Satisfactory	Less than satisfactory	Poor	Excellent	Very Good	Good	Satisfactory	Poor	Very High	High	Medium	Normal	Low	
1.1 Appropriateness of Mission																							
1.1.1 The mission for the program should be consistent with the mission of the institution.	🟥					🟥					🟧						🟥	🟥					**
1.1.2: The mission should establish directions for the development of the program that are appropriate for a program of its type and the needs of students in Saudi Arabia.		🟩					🟥					🟥					🟥		🟧				*
1.1.3: The mission should be consistent with Islamic beliefs and values.		🟩			🟨					🟨						🟧					🟩		***
1.1.4: The mission should be explained to its stakeholders in an accompanying written statement.		🟩		🟩					🟩						🟨							🟩	****

Fig. 7. Smart Rubric-based Screenshot for Evaluating and Prioritizing Academic Practices NCAAA 2009 Standards.



Standard 1: MISSION AND GOALS																											
Criterion Number and Description	Importance Level of Criterion		Extent of availability of elements and components of the criterion				Quality level of application for each element					Regularity of application and assessment, and availability of evidence					Continuous improvement and level of results in the light of indicators and benchmarks				Excellence and creativity in practices of the elements of the criterion	Priority for improvement based on the weakness and strength					Criterion Evaluation
	High	Normal	All Elements Available	Most of the Elements Available	Few Available Elements	No Available Elements	Applied at Distinct Level	Applied at Perfect Level	Applied at Good Level	Applied at Low Level	Not Applied at all or Applied at a Very Low Level	Applied Regularly / Regular Effective, Excellent Assessment / Comprehensive, Cumulative	Applied Regularly / Regular and Effective Assessment / Sufficient and Varied Evidences	Applied Regularly / Regular and Effective Assessment / Sufficient Evidences	Applied Irregularly / (No Assessment (or) Irregular Assessment) / Insufficient Evidences	Rarely Applied	Regular Improvement Procedures and Distinct Results Compared To Other	Regular Improvement Procedures and Higher Results Compared to Previous	Regular Improvement Procedures and Good Results	Limited Improvement Procedures	Creativity in the Practices of the Elements of the Criterion.	Very High	High	Medium	Normal	Low	
1.1 Appropriateness of Mission																											
The institution's mission defines the purpose of its existence; is consistent with its nature, the needs of the community, and the national trends; and is	🟥					🟥					🟥					🟥				🟥		🟥					1
The institution's goals are linked to its mission, clear and realistic. *		🟩		🟩					🟩						🟨				🟨				🟫				2
The institution's mission and goals are formally approved by the governing body and are widely publicized.		🟩	🟩						🟩					🟩				🟩							🟩		3
The mission guides all of the institution's operations (e.g. planning, decision-making, resource allocation, academic program development). *	🟩		🟩				🟩					🟩					🟩				🟩					🟩	5

Fig. 8. Smart rubric-based Screenshot for Evaluating and Prioritizing Academic Practices NCAAA 2018 Standards.



Criterion / Practice Description	Priority for improvement based on the criterion/ practice performance evaluation					STARS
	Very High	High	Medium	Normal	Low	
The mission for the program should be consistent with the mission of the institution.	🟥					**
The program has the sufficient number of qualified staff to perform its administrative, professional and technical tasks, and they have defined tasks and authorities. *	🟥					**
The program forms an advisory committee, comprised of members of professionals and experts in the program specialization, to contribute to its evaluation, development, and performance improvement. *	🟥					**
The graduate attributes and learning outcomes are consistent with the requirements of the Saudi Arabia Qualifications Framework (SAQF) and with academic, professional, and labor market requirements.	🟥					**
The mission should establish directions for the development of the program that are appropriate for a program of its type and the needs of students in Saudi Arabia.		🟫				*
The program study plan considers the adequate requirements for the different tracks (if any) in accordance with international practices and similar programs.		🟫				**
Teaching and learning strategies and assessment methods are aligned with the intended learning outcomes at the program and course levels.		🟫				**
Teaching and learning strategies are student-centered and encourage active learning.		🟫				**

Fig. 9. Suggested Prioritized Action Plan According to the Criterion/ Practice Performance Evaluation.

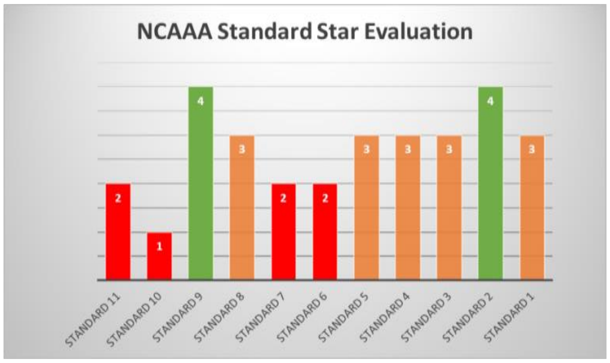


Fig. 10. A Comparison of the NCAAA 2009 Standards Performance Evaluation.

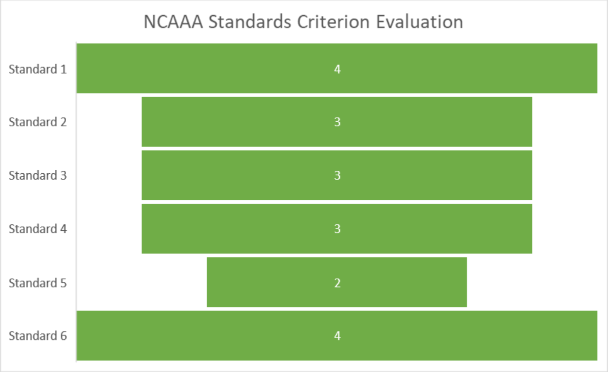


Fig. 11. A Comparison of the NCAAA 2018 Standards Performance Evaluation.

Moreover, the smart rubric can provide analysis of the improvement priority for the standards as shown in Fig. 12 which will support the institutions to focus more in the improvement actions on the standards that need more priority for improvement.

The smart rubric can also provide a comparison of the performance evaluation at the criterion/ practice level as shown in Fig. 13. Thus, an action plan can be implemented at the criterion/ practice level.

Moreover, the smart rubric can provide analysis of the institution / program total quality performance status based on the criterion/ practice improvement priority as shown in Fig. 14.

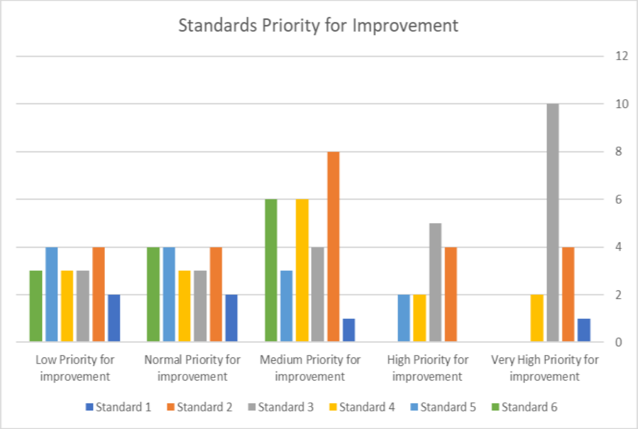


Fig. 12. An Analysis of the Improvement Priority for the Standards.

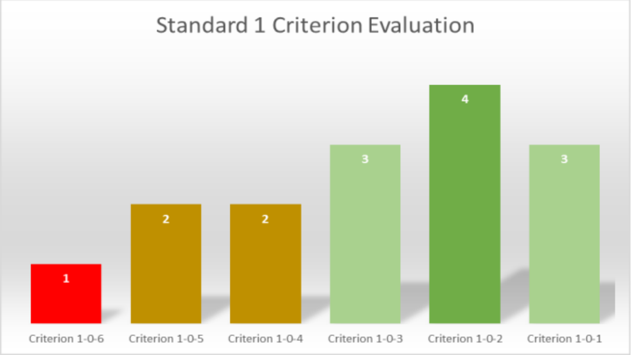


Fig. 13. A Comparison of the NCAAA 2018 Criterion Performance Evaluation.

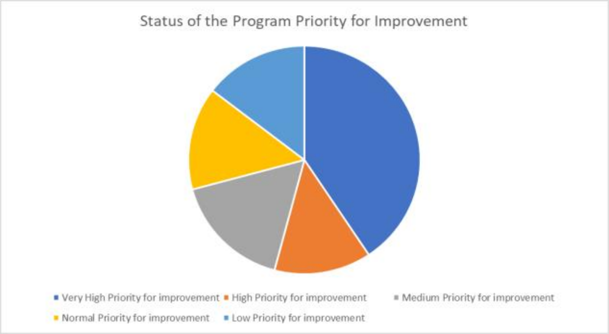


Fig. 14. An Analysis of the Institution / Program Total Quality Performance Status based on the Criterion/ Practice Improvement Priority.

Thus, if the analysis shows that many criterion/ practices needs very high or high priority for improvement, it means the institution / program total quality performance is low. On the other hand, if the analysis shows that many criterion/ practices needs normal or low priority for improvement, that means the institution /program total quality performance is high. Based on that, the smart rubric can provide a specific percentage about the institution's / program's total quality performance status as shown in Fig. 15.

In addition, the smart rubric can provide comparison of standards performance improvement compared to previous self-assessments shown in Fig. 16 to help institution / program to analyze the enhancement actions trend across different assessment cycles.

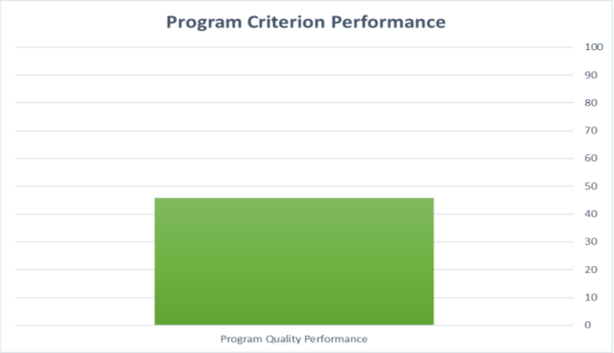


Fig. 15. A Specific Percentage about Institution / Program Total Quality Performance Status.

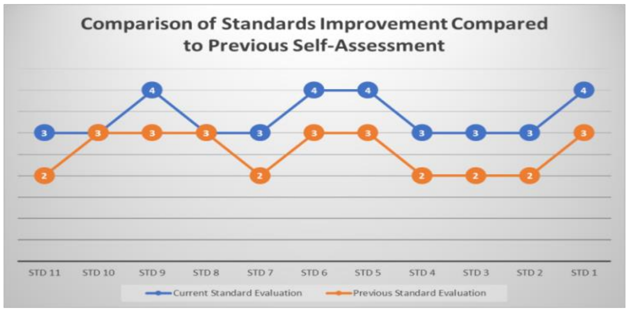


Fig. 16. A Comparison of Standards Performance Improvement Compared to Previous Self-Assessment Cycle.

V. Conclusion

NCAAA standards in Kingdom of Saudi Arabia aim to prepare highly qualified graduates to meet the new challenges causes by the impact of free-market economy, globalization, and knowledge economy. However, implementing NCAAA standards without supportive systems has been found to be a very complex task. In this paper, we have described the development of a very sustainable and efficient smart rubric-based systematic model for evaluating and prioritizing NCAAA criterions/ practices and developing an accurate quality action plans based on the criterions/practices evaluation. The implementation of the proposed smart rubric-based systematic model demonstrates a high degree of validity, usefulness, accuracy for developing an implementation action plan. Moreover, reduces the time and efforts for evaluating NCAAA criterions/ practices by auto-evaluating, auto-calculating the star/ point, and auto-prioritizing and suggesting priority for improvement of criterion/ practice. Furthermore, the proposed smart rubric-based systematic model supports the academic institution's/ program's decision making by providing analysis of the standards improvement priority, analysis of the performance evaluation at the criterion/ practice level, analysis of the total quality performance status, analysis of standards performance improvement compared to different assessment cycles, and provides a specific percentage of the total quality performance status. Therefore, Saudi higher educational institution and programs can implement accurate action plans that will lead to enhance the quality of education and to obtain NCAAA accreditation.

References

[1] I. Chirikov, How global competition is changing universities: three theoretical perspectives, 2016.

[2] J. Zajda, Globalisation and Its Impact on Education and Policy, In: Zajda J, editor, Second International Handbook on Globalisation, Education and Policy Research, Dordrecht: Springer Netherlands, p. 105-25, 2015.

[3] J. Zajda, V. Rust, Current Research Trends in Globalisation and Neo-liberalism in Higher Education, In: Zajda J, Rust V, editors, Globalisation and Higher Education Reforms, Cham: Springer International Publishing, p. 1-19, 2016.

[4] N. Salma, "Impact of Globalization on Higher Education in Pakistan: Challenges and Opportunities" International Journal of Innovation in Teaching and Learning (IJITL), 2019.

[5] National Commission for Academic Accreditation and Assessment–NCAAA, Standards for Quality Assurance and Accreditation of Higher Education Institutions, Saudi Ministry of Higher Eductuion, 2009.

[6] National Commission for Academic Accreditation and Assessment–NCAAA, Standards for Quality Assurance and Accreditation of Higher Education Programs, Saudi Ministry of Higher Educ., Riyadh, KSA; http://ncaaa.org.sa, . 2009.

[7] National Center for Academic Accreditation and Assessment–NCAAA. Standards for Institutional Accreditation, Saudi Ministry of Higher Education., Riyadh, KSA; http://ncaaa.org.sa, 2018.

[8] National Center for Academic Accreditation and Assessment–NCAAA. Standards for Program Accreditation, Saudi Ministry of Higher Educ., Riyadh, KSA; http://ncaaa.org.sa, 2018.

[9] National Center for Academic Accreditation and Assessment–NCAAA. Standards for Postgraduate Program Accreditation, Saudi Ministry of Higher Educ., Riyadh, KSA; http://ncaaa.org.sa, 2018.

[10] Moskal, B. M., & Leydens, J. AScoring rubric development: Validity and reliability. Practical assessment, research & evaluation, 2000. 7(10), 71-81.

[11] National Commission on Academic Accreditation and Assessment–NCAAA. Self Evaluation Scales for Higher Education Institutions, Saudi Ministry of Higher Educ., Riyadh, KSA; http://ncaaa.org.sa , 2009.

[12] National Commission on Academic Accreditation and Assessment–NCAAA. Self-Evaluation Scales for Higher Education Programs, Saudi Ministry of Higher Educ., Riyadh, KSA; http://ncaaa.org.sa, 2009.

[13] National Center for Academic Accreditation and Assessment–NCAAA. Self-Evaluation Scales for Higher Education Institutions., Riyadh, KSA; http://ncaaa.org.sa, 2018.

[14] National Center for Academic Accreditation and Assessment–NCAAA. Self-Evaluation Scales for Higher Education Programs., Riyadh, KSA; http://ncaaa.org.sa National Center for Academic Accreditation and Assessment–NCAAA, 2018.

[15] Deanship of development and quality, Itqan 2020 KSU–QMS (Quality Management System) an Introductory synopsis, 4th Edition, Kinh Saud Unvirsoty, 2018.

[16] D. Khalid Al-Harthi, I. Albidewi, "Balanced scorecard automation system in higher education institutions", European Journal of Computer Science and Information Technology, Vol.5, No.4, pp.25-49, 2017.

[17] National Center for Academic Accreditation and Assessment–NCAAA. Electronic Accreditation System. Retrieved from https://daman.ncaaa.org.sa/ar/login, 2018.

Modeling of the Vegetable Oil Blends Composition

Olga S. Voskanyan[1]
Department "Technology of Products from Plant Materials, Perfumes and Cosmetics", K.G. Razumovsky Moscow State University of Technologies and Management (The First Cossack University), Moscow, Russian Federation

Igor A. Nikitin[2]
Department "Technology of Grain Processing, Bakery, Pasta and Confectionery Industries"
K.G. Razumovsky Moscow State University of Technologies and Management (The First Cossack University)
Moscow, Russian Federation

Marina A. Nikitina[3]
V.M. Gorbatov Federal Research Center for Food Systems
Moscow
Russian Federation

Maria V. Klokonos[4]
Department "Technology of Grain Processing, Bakery, Pasta and Confectionery Industries", K.G. Razumovsky Moscow State University of Technologies and Management (The First Cossack University), Moscow, Russian Federation

Daria A. Guseva[5]
Department "Technology of Grain Processing, Bakery, Pasta and Confectionery Industries"
K.G. Razumovsky Moscow State University of Technologies and Management (The First Cossack University)
Moscow, Russian Federation

Igor V. Zavalishin[6]
K.G. Razumovsky Moscow State University of Technologies and Management (The First Cossack University)
Moscow, Russian Federation

***Abstract*—The article presents a computer modeling of blends of vegetable oils for treatment-and-prophylactic and healthy nutrition. To solve this problem, based on biomedical requirements, models of vegetable oil blends have been developed, taking into account the required chemical composition, mass fractions of the main components of the product, structural correlations of the biological value of blends according to the criterion of fatty acid compliance (omega-6 to omega-3). The problem was solved by the method of complete enumeration (brute force), which belongs to the class of methods for finding a solution by exhausting all sorts of options. An automated research system has been developed and implemented to model the composition of mixtures of vegetable oils in accordance with a given target function of the ratio of omega-6: omega-3 polyunsaturated fatty acids (PUFAs). The use of an automated system allowed us to model prescription formulations of blends of oils, taking into account the constraints set and a given goal function. In this study, three alternatives were obtained for a blend composition with omega-6: omega-3 in the first and second variant 5: 1, and in the third - 10: 1, which makes it possible to use them for healthy and therapeutic nutrition.**

Keywords*—Modeling; brute force; vegetable oil blends; omega-6; omega-3

I. Introduction

According to the scientific and technical policy in the field of healthy and safe nutrition, modern food products should not cause damage to human health, should satisfy physiological needs, as well as perform therapeutic and preventive functions. Ensuring the fulfillment of these tasks is possible only through the development of a direction related to the "design" of new food products based on the fulfillment of criteria for food adequacy of a given level, and the solution to the problem of optimizing the composition of multicomponent food products can be described mathematically [1]. Currently, food design principles are based on modern knowledge in the field of physiology, molecular medicine and food chemistry, and they are based on the concept of a balanced diet, which determines the correlative relationship between food absorption and the degree of balance of its chemical composition.

Therefore, creation of new generation fatty products of functional and specialized purposes, balanced in their fatty acid composition, for the oil and fat industry is one of the promising areas of innovative development of the food industry [2]. Based on the variation of ratios of the known vegetable oils in the blends, you can create various products with desired functional and therapeutic properties [3, 4].

The aim of the work was to develop an integrated approach to creating compounding blends of vegetable oils, optimized by the composition of polyunsaturated fatty acids for healthy and therapeutic and preventive nutrition, in particular with a given ratio of omega-6 / omega-3=5/1 polyunsaturated fatty acids using the "brute force" method.

II. Formulation and Analysis of the Problem of Development of Oil Blends with Balanced Fatty ACID Composition

Vegetable oils are vital products that have a large and diverse use in the human diet [5]. Compared with carbohydrates and proteins, lipids have a higher nutritional and energy value, so nutritionists attach great importance to the issue of increasing the proportion of vegetable oils used in food, because they have a specific physiological effect due to the content of essential polyunsaturated acids positively affecting the human body.

Excess intake of fats rich in saturated fatty acids contributes to the development of atherosclerosis and coronary heart disease, obesity, gallstone disease. At the same time, modern man lacks polyunsaturated fatty acids (PUFAs), the source of which is mainly vegetable oils [6].

Omega-3 and omega-6 PUFA are essential fatty acids that the human body cannot synthesize on its own and can only be obtained from food. However, the effect of these fatty acids on the human body is manifested quite differently [7]. Since Omega-3 and Omega-6 PUFAs compete for the same delta-6-desaturase enzyme, with the help of which they are converted to longer-chain acids, the ratio of these fatty acids has a significant effect on the ratio of eicasoids that act as tissue hormones in the body , and, therefore, significantly affects the metabolic processes in the body.

According to numerous studies, it has been established that these fatty acids should come from food in the ratio of omega-6 / omega-3, equal to 8-9 / 1 for healthy people, and in the ratio 5/1 for therapeutic and preventive nutrition, while modern man eats omega-3 with food 5-6 times less [8]. In this regard, the development of blends containing the optimal ratio of omega-6 / omega-3 fatty acids based on various vegetable oils for therapeutic and prophylactic nutrition is an urgent task, and the creation of software tools to automate this process will contribute to its operational solution.

In the process of developing the composition of blends, it is advisable to apply computer modeling methods that allow to determine functional properties more accurately and establish the optimal percentage of ingredients that make up the blends [9, 10].

For the preparation of new science-based blends of vegetable oils used in therapeutic and preventive nutrition, methods of mathematical modeling of the selection of the composition, in particular the brute force method, have recently been used.

To solve the problem, it is first necessary to analyze and select vegetable oils with a high content of omega-3 PUFA. Next, you need to choose a blend composition that would meet the medical and biological requirements and restrictions. To do this, you can use the process of optimizing the composition of new blends of vegetable oils on the ratio of omega-6 PUFA: omega-3, which is 10: 1 for healthy people, and 5: 1 for therapeutic and preventive nutrition. For this purpose, it is necessary to model the component composition of vegetable oil blends with the content of optimal fatty acid composition for therapeutic and preventive nutrition (omega-6: omega-3 = 5: 1) that meets the relevant requirements for physical and chemical quality indicators, as well as having organoleptic properties, satisfying tastes of most consumers.

The solution of the task consists of three stages:

- Input of the initial data on the components of blends of oils and PUFAs (omega-6 and omega-3 acids);
- The choice of the objective function by the ratio of PUFA;
- Carrying out the necessary calculations.

TABLE. I. FATTY ACID COMPOSITION OF VEGETABLE OILS

Name of PUFA	**The content of PUFA in vegetable oil,% of the total**					
	Sunflower oil	***Mustard oil***	***Pumpkin oil***	***Linseed oil***	***Wheat germ oil***	***Milk Thistle oil***
Linoleic (omega-6)	50.90	17.80	53.10	15.89	57.00	55.60
Arachidonic (omega-6)	0.50	0.00	0.30	0.00	0.00	2.00
б-Linolenic (omega-3)	0.30	5.60	8.00	60.00	5.60	3.00

The selection of components of the composition of new oil blends was carried out on the basis of the analysis of literature data [11]. Characteristics of the raw ingredients of blends of oils for therapeutic and prophylactic nutrition are presented in Table I.

Analyzing Table I, first of all it should be noted that sunflower and pumpkin oils, as well as wheat germ oil and milk Thistle oil contain significant amounts of linoleic acid, and at the same time, arachidonic acid is practically absent in them. Considering the group of omega-3 acids, it is necessary to highlight linseed oil, which contains 60.0% α-linolenic acid [12]. The rest of the studied oils are characterized by a low content of this acid-ranging from 0.3% (in sunflower oil) to 8.0% (in pumpkin oil).

Such an incomplete presence of acids of the omega-6 and omega-3 groups poses the problem of optimizing the fatty acid composition for the development of blends of vegetable oils for therapeutic and prophylactic purposes [13]. At the same time, they are prepared on the basis of optimal ratios of linoleic and α-linolenic acids, since other acids are contained in the blend oils in an insignificant amount and have no significant effect on the balance of acids [14, 15].

The solution of the problem is possible by modeling the composition of the blends using the exhaustive “brute force” method with a numerical value of 5: 1 as the objective function (the ratio of omega-6 to omega-3 as 5: 1) [16].

III. ALGORITHM OF SCIENTIFIC RESEARCH SYSTEM WORK. MATHEMATICAL FORMALIZATION

The task was implemented by the method of complete enumeration with a given accuracy of 1%.

A complete search (or the “brute force” method) is a method for solving mathematical problems. Refers to a class of methods for finding solutions by exhaustion of all sorts of variants [17]. The complexity of the method depends on the number of all possible solutions to the problem. The problem considered in the framework of our study belongs to the NP class, therefore it can be solved by this method.

The algorithm of the proposed method is presented in Fig. 1.

Pointers allow you to restore the path back after finding the recipe composition at the i-th iteration step. In step 5 of the presented algorithm, the recipe composition is placed at the end of the Open list.

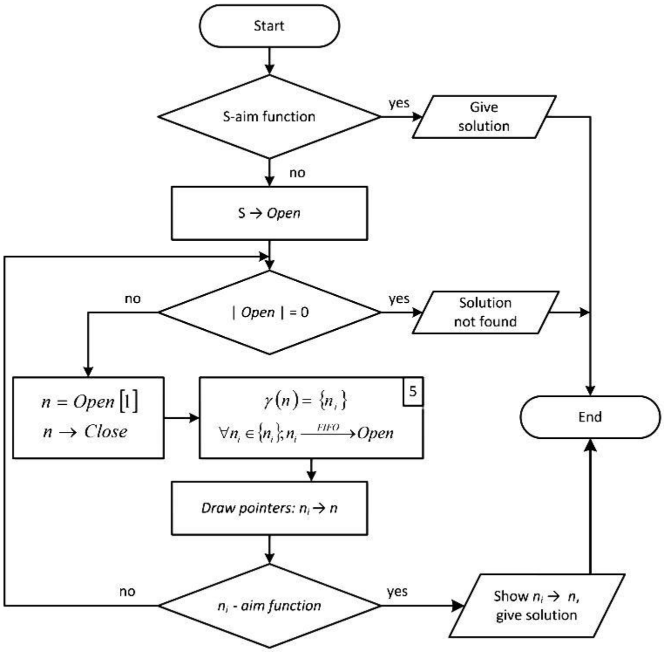


Fig. 1. Block Diagram of the "Brute Force" Method, where "Open" is a List of Prescription Compositions to which the Operator γ has not yet been Applied (Optimization Criterion and Constraints); "Close"-a List of Prescription Compositions to which the γ Operator has Already been Applied.

To solve the problem of optimizing the composition of plant blends in various formulations and combinations of non-linear criteria and restrictions we perform simulation modeling with playing out all possible combinations of the initial components of the formulation, followed by checking the restrictions and calculating the criteria using the algorithm of the brute force method.

The search for the optimal blend of vegetable oils from n-components begins with the choice of the mass fraction of the first component X_1 and the search of other values that mimic all combinatorial versions of the n-component product formulation. Further, if X_j satisfies the boundary conditions, then the next acceptable ratio is memorized and the cycle goes to the next relation for j = j + 1-th component. When $X_j < min_j$, its fractional ratio with subsequent components increases by the step size, and in the case of $X_j > max_j$, it returns to the initial jth ratio and continues the previous cycle for j = j – 1th component with a starting value.

After selecting the specified ratios of all the components of the recipe so that the mass fractions in the total amount to one, the parametric and balance constraints are checked with an accumulation of permissible options evaluated by the objective function and the choice of the best alternative solution.

The algorithm for the sequential formation of the ingredients included in the recipe allows you to find the optimal model as follows:

1) The first combination of recipes that satisfies all the boundary and balance constraints is considered, and its value of the criterion $F_1(x)$ is calculated. This recipe is considered optimal and its parameters are captured.

2) The second composition of the formulation that satisfies all the boundary and balance constraints is considered, and its value of the criterion $F_2(x)$ is calculated. If $F_2(x) < F_1(x)$, then the second composition is optimal, its recipe is captured, and the recipe of the previous model is deleted. If $F_1(x) <= F_2(x)$, then the previous formulation remains optimal and the next model is considered.

3) The process of reviewing new formulation compositions and comparing the quality of the new formulation and the best in the previous step continues until all possible combinations of the formulation compositions are considered.

As a source of data, a list of raw materials was used (Table I) for modeling the composition of oil blends. Omega-6 and omega-3 values in 1 gram of lipid were used as the coefficients of the objective function. In addition, the study took into account the boundary data for each component in the composition of the blends.

When developing options, the choice of the minimum and maximum dosage of the components in the blends was determined experimentally.

Analysis of the results of the research allowed us to establish the permissible limits for the variation of the boundary conditions for vegetable oils in the blends (Table II).

The analysis of the compatibility of the aromatic characteristics of the vegetable oils under consideration served to compile three blend compositions:

1) blend: mustard oil, pumpkin oil and milk thistle oil;
2) blend: sunflower oil, linseed oil and milk thistle oil;
3) blend: sunflower oil, wheat germ oil and milk thistle oil.

The mathematical formalization of the task in generalized form is as follows:

objective function

$$\Phi(X) = \frac{\omega_6}{\omega_3} \rightarrow k$$

under restrictions

$$\omega_i = \sum_{i=1}^{n} \omega_i x_i$$

$$\sum_{i=1}^{n} x_i = 1, x_i > 0$$

where ω_i is the content of PUFAs of the i-th omega family (i = 3 - oils of the omega-3 family, i = 6 - oils of the omega-6 family);

x_i is the mass fraction of the i-th vegetable oil in the blend;

k is the ratio to which the ratio of omega-6 (ω6) to omega-3 (ω3) should tend, in our case 5:1.

TABLE. II. PERMISSIBLE LIMITS OF VARIATION OF THE BOUNDARY CONDITIONS FOR VEGETABLE OILS IN THE BLENDS' COMPOSITION

Oil name	min, %	max, %
Sunflower oil	0	60
Mustard oil	0	60
Linseed oil	0	30
Milk Thistle oil	0	60
Wheat germ oil	0	90
Pumpkin oil	0	40

For our particular case, mathematical models are presented in three alternative versions as follows:

1) *For the first blend*

$$\omega6 / \omega3 = 5/1$$

$$\omega_6 = 17.8X_1 + 53.1X_2 + 55.6X_3$$

$$\omega_6 = 5.6X_1 + 8.0X_2 + 3.0X_3$$

2) *For the second blend*

$$\omega6 / \omega3 = 5/1$$

$$\omega_6 = 51.5X_1 + 15.89X_2 + 55.6X_3$$

$$\omega_6 = 0.3X_1 + 60.0X_2 + 3.0X_3$$

3) *For the thirtd blend*

$$\omega6 / \omega3 = 5/1$$

$$\omega_6 = 51.5X_1 + 57.0X_2 + 55.6X_3$$

$$\omega_6 = 0.3X_1 + 5.6X_2 + 3.0X_3$$

IV. IMPLEMENTATION OF THE PROGRAM OF THE AUTOMATED SCIENTIFIC RESEARCH SYSTEM RESULTS AND DISCUSSION

Computer modeling using the "brute force" method for the NP problem class allows you to get a complete set of solutions, to make the choice of the optimal solution from the resulting set to achieve the specified constraints and goals.

The method used allows to:

1) Get a complete set of solutions to the problem and suitability for obtaining any other blends and their relationship.

2) Get a sufficient amount of data (with an accuracy of 1%) to select optimal blends from a variety of computer solutions.

The information basis of a computer program is a database of the chemical composition of vegetable oils. It includes characteristics such as fat content, saturated fatty acids (EFA), monounsaturated fatty acids (MUFA), PUFA, and vitamin E (alpha-tocopherol). The database provides for the search of vegetable oils for a specific attribute. As an example in Fig. 2 information on the content of omega-6 and omega-3 in oils is provided.

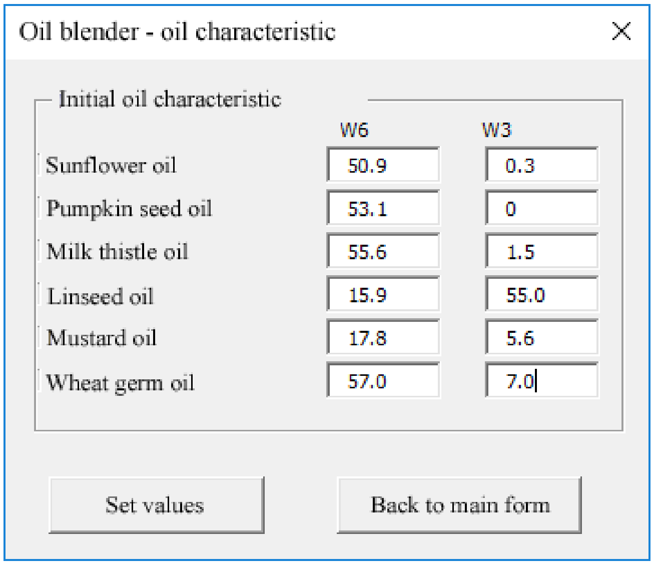
Restrictions on the maximum concentration of oil in a blend, as well as the numerical value to which the objective function should strive are set in the program window (Fig. 3).

The computer system provides the ability to export the results in MS Excel. In Fig. 4, one of the alternative solutions for the operation of an automated research system is presented.

As a result of the work of the automated research system, about 300 blends of vegetable oils were generated. A detailed and reasonable analysis of the results of the calculated data led to the following conclusions:

For the first blend (mustard oil, pumpkin oil and milk thistle oil) and the second blend (sunflower oil, linseed oil and milk thistle oil) the ratio of omega-6: omega-3 can reach values of 5: 1 and above. Consequently, this blend can be used for therapeutic and prophylactic nutrition.

For the third blend (sunflower oil, wheat germ oil and milk thistle oil), the omega-6 / omega-3 ratio can reach values of 10: 1 and above. Therefore, the third blend can only be used for healthy eating.

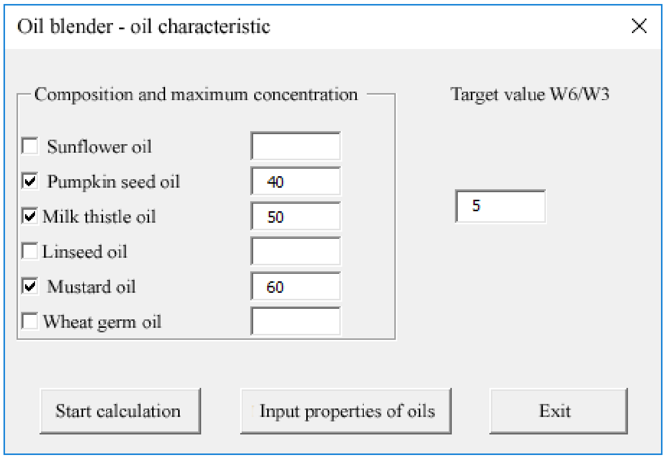


Oil blender - oil characteristic

Initial oil characteristic

	W6	W3
Sunflower oil	50.9	0.3
Pumpkin seed oil	53.1	0
Milk thistle oil	55.6	1.5
Linseed oil	15.9	55.0
Mustard oil	17.8	5.6
Wheat germ oil	57.0	7.0

Fig. 2. Fragment of the Russian Version of the Database "Characteristics of Oils on the Composition of Omega-6 and Omega-3 PUFA".



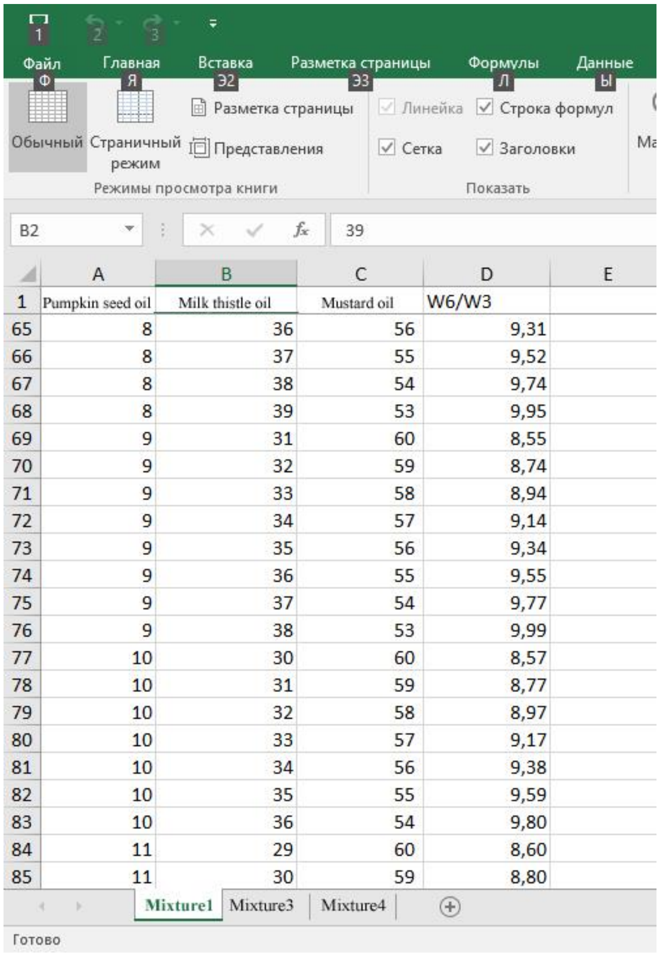
Oil blender - oil characteristic

Composition and maximum concentration

□ Sunflower oil	
☑ Pumpkin seed oil	40
☑ Milk thistle oil	50
□ Linseed oil	
☑ Mustard oil	60
□ Wheat germ oil	

Target value W6/W3: 5

Fig. 3. Input of Initial Data into the Program.



	A	B	C	D	E
1	Pumpkin seed oil	Milk thistle oil	Mustard oil	W6/W3	
65	8	36	56	9,31	
66	8	37	55	9,52	
67	8	38	54	9,74	
68	8	39	53	9,95	
69	9	31	60	8,55	
70	9	32	59	8,74	
71	9	33	58	8,94	
72	9	34	57	9,14	
73	9	35	56	9,34	
74	9	36	55	9,55	
75	9	37	54	9,77	
76	9	38	53	9,99	
77	10	30	60	8,57	
78	10	31	59	8,77	
79	10	32	58	8,97	
80	10	33	57	9,17	
81	10	34	56	9,38	
82	10	35	55	9,59	
83	10	36	54	9,80	
84	11	29	60	8,60	
85	11	30	59	8,80	

Fig. 4. Data Export to MS Excel. the Results of the Automated Research System for the №1 Blend.

TABLE. III. COMPONENT COMPOSITIONS OF BLENDS OF VEGETABLE OILS AND THE RATIO OF OMEGA-6 / OMEGA-3 PUFA OBTAINED FOR THEM

	Blend 1	**Blend 2**	**Blend 3**
Sunflower oil	-	60	2
Mustard oil	58.27	-	-
Pumpkin oil	40	-	-
Linseed oil	-	14.24	-
Wheat germ oil	-	-	90
Milk Thistle oil	1.73	25.76	8
Omega-6 / omega-3 ratio	5	5	10

The results of mathematical design and structural optimization of the composition of vegetable blends are shown in Table III.

Creating a blend of vegetable oils that combine high biological value, good aromatics (organoleptics) and compensate the lack of omega-6 and omega-3 requires: selection of components of the formulation; knowledge of structural relationship and principles of food combinatorics. All this knowledge was taken into account when creating an automated system. The use of the system allows to model prescription formulations of blends of oils, taking into account the prescripted limits and the specified goal. In the study three alternative blend compositions with an omega-6: omega-3 ratio in the first and second variant 5: 1 were obtained from the proposed range of vegetable oils, and 10: 1 in the third, which makes it possible to use them for healthy and healthy-preventive nutrition.

V. CONCLUSION

As a result of the research conducted on the basis of biomedical requirements, the choice of potential ingredients for blending vegetable oils was justified; compositions of vegetable oils of a given quality and properties are designed. Mathematical models of designed blends were developed in the form of a set of criteria and restrictions. One of the main limitations in the problem being solved was the level of maximum concentration of ingredients, so that in the resulting blend, none of the plant components would prevail over the others and, accordingly, could not negatively affect the organoleptic properties of the finished composition (the target function was the omega-6 to omega-3 ratio). As a result, with the help of information technology that implement the methods of mathematical programming, namely, the method of complete enumeration (brute force), three optimal alternative versions of the composition of vegetable oils were designed.

As subsequent tasks for the study, it would be advisable to develop an integrated large-scale updated database of the fatty acid composition of various vegetable and animal oils for wider possibilities in the design of blends. It would also be important to be able to use different optimization criteria (across the entire spectrum of the distinctive properties of vegetable oils) when making blends depending on the tasks set.

ACKNOWLEDGMENT

The work was carried out with the financial support of the Russian Foundation for Basic Research in the framework of grant No. 19-016-00049 “Development of a software system for automated calculation of the composition of flour mixtures of functional and dietary nutrition”.

REFERENCES

[1] Ivashkin, Yu.A. Information Technologies for Food Design / Yu.A. Ivashkin, S.B. Yudina, M.A. Nikitina, N.G. Azarova // Meat industry. - 2000. - No. 5. - P. 40-41.

[2] Ostryakov, A.N. Blended vegetable oil - a functional food product / A.N. Ostryakov, M.V. Kopylov // Successes in modern science. - 2011. - No. 7. - P. 171-172.

[3] Skoryukin A.N. Technology of production and application of blended products with the optimal composition of PUFA: PhD thesis. - Moscow, 2004. - 24 p.

[4] Tutelyan, V.A., Nechaev A.P., Kochetkova A.A. Functional fatty foods in the nutritional structure / V.A. Tutelian, A.P. Nechaev, A.A. Kochetkova // Oil and fat industry. - 2009. - No. 6. - P. 6–9.

[5] Birbasova A.V. Theoretical and experimental substantiation of the formulations of blended functional oils: PhD thesis. –Krasnodar, 2016. - 24 p.

[6] Levachev M.M. The value of fat in the diet of a healthy and sick person: a guide to dietetics / ed. V.A. Tutelyan, M.A. Samsonova. M., 2002.

[7] Grushina, E.N. On the mechanisms of action of PUFAs on the immune system / E.N. Sruschina, O.K. Mustarina, V.L. Volgarev // Nutrition. - 2003. - №3. - pp.35-39.

[8] Ipatova L.G., Kochetkova A.A., Nechaev A.P., Tutelyan V.A. Fatty foods for a healthy diet. Modern look. M., 2009. 396 p.

[9] Usatnikov, S.V. Evaluation of the effectiveness of using a computer program in creating mixtures of vegetable oils for a healthy diet / S.V. Usatnikov, T.I. Timofeyenko, O.V. Rudenko, S.N. Nikonovich, A.V.

Birbasova, D.A. Ovgareva // Fundamental research. - 2015. - №10 (2).- p. 314-317.

[10] Nikolaeva, S.V. Application of the linear programming method for optimization of mixtures of vegetable oils / S.V. Nikolaev, E.A. Klyushina, E.V. Gruzinov, T. Shlyonskaya // Fat-and-oil industry.-2007. - № 1. - p. 23-24.

[11] Voskanyan, O.S. The current state and development trends in the production of emulsion food / O.S. Voskanyan, V.Kh. Paronyan, T.V. Shlenskaya. - M .: Pishepromizdat, 2003. - 353 p.

[12] Ipatova, O.M. The biological activity of linseed oil as a source of omega-3 alpha-linolenic acid / OM. Ipatova, N.N. Prozorovskaya, V.S. Baranova, D.A. Gusev // Biomedical chemistry. - 2004. - T. 50. - No. 1. - P. 25-43.

[13] Zharinov, A.I. Experimental computer modeling of mayonnaise formulations // A.I. Zharinov, M.Yu. Popova, M.A. Nikitin, V.Yu. Mayauska // Fat-and-oil industry. - 2008. - № 1. - p. 34-37.

[14] Patent RUS No. 2402911from 28.05.2009 N.N. Prozorovskaya, D.A. Guseva, A.V. Shironin, M.A. Sanzhakov, E.G. Tikhonova, O.M. Ipatova Vegetable oil based on a mixture of flax, sesame and milk thistle seeds with a ratio of PUFA omega-3 and omega-6 (1: 1.4: 1.6) and methods for its preparation // patent for invention RUS № 2402911, 2009.

[15] RUS patent No. 2402911ot 28.05.2009 N.N. Prozorovskaya, D.A. Guseva, A.V. Shironin, M.A. Sanzhakov, E.G. Tikhonova, O.M. Ipatova Vegetable oil based on a mixture of flax, sesame and milk thistle seeds with a ratio of PUFA omega-3 and omega-6 (1: 6: 81.6) and methods for its preparation // patent for invention RUS No. 2402912, 2009.

[16] Certificate of state computer program No. 2015660121 O.V. Rudenko, T.I. Timofeyenko, S.N. Nikonovich, A.V. Birbasova, D.A. Ovgareva, S.V. Usatnikov. // Optimization of the composition of fatty acids of vegetable oils for therapeutic and dietary preventive nutrition / Computer № 2015660121 from 07.27.2015.

[17] Thomas H. Cormen et al.,. Introduction to Algorithms. - MIT Press, 2001. - P. 1292. - ISBN 978-0-262-03384-8.

A Machine Learning Approach towards Detecting Dementia based on its Modifiable Risk Factors

Reem Bin-Hezam[1]
Information Systems Department
College of Computer & Information Sciences
Princess Nourah bint Abdulrahman University
Riyadh, Saudi Arabia

Tomas E. Ward[2]
Insight Centre for Data Analytics
Dublin City University
Glasnevin, Dublin 9, Ireland

for the Alzheimer's Disease Neuroimaging Initiative*

***Abstract*—Dementia is considered one of the greatest global health and social care challenges in the 21st century. Fortunately, dementia can be delayed or possibly prevented by changes in lifestyle as dictated through known modifiable risk factors. These risk factors include low education, hypertension, obesity, hearing loss, depression, diabetes, physical inactivity, smoking, and social isolation. Other risk factors are non-modifiable and include aging and genetics. The main goal of this study is to demonstrate how machine learning methods can help predict dementia based on an individual's modifiable risk factors profile. We use publicly available datasets for training algorithms to predict participant' s cognitive state diagnosis, as cognitive normal or mild cognitive impairment or dementia. Several approaches were implemented using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) longitudinal study. The best classification results were obtained using both the Lancet and the Libra risk factor lists via longitudinal datasets, which outperformed cross-sectional baseline datasets. Moreover, using only data of the most recent visits provided even better results than using the complete longitudinal set. A binary classification (dementia vs. non-dementia) yielded approximately 92% accuracy, while the full multi-class prediction performance yielded to a 77% accuracy using logistic regression, followed by random forest with 92% and 70% respectively. The results demonstrate the utility of machine learning in the prediction of cognitive impairment based on modifiable risk factors and may encourage interventions to reduce the prevalence or severity of the condition in large populations.**

Keywords—Machine learning; classification; data mining; data preparation; dementia; modifiable risk factors

I. Introduction

Dementia presents enormous global health and social challenges. Currently, there are around 47 million people with dementia worldwide, and that number is expected to triple by 2050. Dementia occurs mainly in people older than 65 years [1]. The aging population worldwide is almost certainly part of the reason behind this increase, especially in low- and middle-income countries.

Dementia is a collection of symptoms of cognitive defects, which could be delayed or possibly prevented by eliminating certain modifiable risk factors associated with the condition. However, few researches used machine learning approaches to detect dementia based on its modifiable risk factors, while most of the previous researches used machine learning to detect dementia based on imaging data or non-modifiable factors such as genetics. Although these methods are useful in diagnosing dementia, they may not be as much useful in term of delaying or preventing dementia as there is nothing that could be modified.

This study aims to use a machine learning (ML) approach to classify the cognitive state and detect dementia based only on the modifiable risk factors. The main research objective is to determine to what extent it is possible to predict dementia based on an individual's modifiable risk factors profile.

The analysis of this research is applied to data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) longitudinal study, using modifiable risk factors lists that have been already defined by the *Lancet* commission and the *Libra* index. As far as known, no previous work has explored *Lancet*, and *Libra* lists of modifiable risk factors on the ADNI dataset using a machine learning approach.

Moreover, being able to accurately detect dementia based only on its modifiable risk factors would make it possible not only to predict dementia but also to target its risk factors. This will be useful in term of trying to delay or prevent the disease by eliminating these factors as possible.

The remaining of this paper is structured as follows. Section II provides background on the domain and some related work. The methodology applied in this research is described in Section III. Moreover, the experiment and results are provided in Section IV. Finally, the conclusion of the research and its future work is provided in Section V.

II. Background

Dementia is described as a collection of symptoms related to cognitive deficits and is not considered one single disease. In the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [2], dementia is listed under Major Neurocognitive Disorder (NCD), and is defined by the following:

- There is evidence of a substantial cognitive decline in one or more cognitive domains.
- The cognitive deficits interfere with independence in everyday activities, are not exclusively in the context of a delirium, and are not mainly attributable to another mental disorder.

* Membership of the Alzheimer's Disease Neuroimaging Initiative can be found in the Acknowledgment section.

Risk factors for dementia can be either modifiable or non-modifiable [3]. Fortunately, dementia could be delayed or possibly prevented by eliminating some of its modifiable risk factors [4].

A. Dementia Risk Factors

The *Lancet* Commission study found that around 35% of dementia risk factors are potentially modifiable [1]. These risk factors include less education, hypertension, obesity, hearing loss, depression, diabetes, physical inactivity, smoking, and social isolation. Although the impact of these factors varies at different life stages, eliminating them at any stage would be beneficial. Moreover, studies recommend active treatment and intervention of modifiable dementia risk factors, which would potentially delay or prevent 30% of dementia cases [1], [4].

On the other hand, completely eliminating the apolipoprotein E (APOE) ε4 allele, which is considered the major genetic risk factor of dementia, could reduce its incidence by 7% [1]. However, this and all other genetic factors are considered to be non-modifiable. Besides genetics, other non-modifiable risk factors include age and gender.

A common method to calculate dementia risk based on its risk factors is by using the Lifestyle for Brain Health (LIBRA) index [5], [3], [6], which is calculated by the Innovative Midlife Intervention for Dementia Deterrence (In-MINDD) project [7].

Table I listed the modifiable risk factors defined by both *Lancet* commission and *Libra* index.

The National Academy of Medicine committee [8] also identified cognitive training, blood pressure management for people with hypertension, and increased physical activities as three main classes of dementia intervention.

Alzheimer's disease (AD) is the most common type of dementia. The next most common type is vascular dementia (VaD), followed by dementia with Lewy bodies. Frontotemporal degeneration and dementia associated with brain injury, infections, and alcohol abuse are less common types of dementia [1].

TABLE. I. DEMENTIA RISK FACTORS AND DIAGNOSIS ATTRIBUTES

#	Risk Factor	Lancet List	Libra List
1	Low Education	●	●
2	Hypertension	●	●
3	Obesity	●	●
4	Smoking	●	●
5	Depression	●	●
6	Diabetes	●	●
7	Physical inactivity	●	●
8	Hearing loss	●	
9	Social isolation	●	
10	Cognitive inactivity		●
11	Chronic Heart disease		●
12	Alcohol use		●
13	Chronic Kidney disease		●

Tariq and Barber [9] suggested dementia prevention by targeting vascular modifiable risk factors, as these two types are often co-existing in the brain and share some common modifiable risk factors.

B. Current Approaches used in Detecting Dementia Risk Factors

Many studies have aimed to predict an early diagnosis of dementia through magnetic resonance imaging (MRI) and genetic variables [10]. However, these measurements are expensive and not always available.

Most of the research that has used machine learning applied classification methods from MRI data to classify or predict a diagnosis of different cognitive diseases and states [11], [12], [13].

On the other hand, only a few studies have used machine learning techniques to determine risk factors associated with dementia or one of its major causes (i.e., Alzheimer's disease) [11]. Some of the studies combined modifiable and non-modifiable risk factors in order to reach a higher level of accuracy.

Most of the available research used large cohort studies and a population-based perspective to determine associated risk factors [14], while some used statistical analysis to provide a ranked risk-factor index [3], [6].

Two main studies used machine learning techniques to detect dementia's risk factors and predict dementia risk accordingly [15], [16]. Both studies applied their analysis to one longitudinal cohort study with a relatively small size (i.e., 840 and 746 subjects respectively).

O'Donoghue, et al. [15] applied a non-linear dementia survival prediction model with a multilayer perceptron (MLP), which is an artificial neural network (ANN), and used both modifiable and non-modifiable risk factors defined in the In-MINDD project [5]. They also examined the hidden layers to extract different clusters of risk factors and explore different interactions between them. Due to a class imbalance of the MAAS dataset, their models were able to predict survival better than predicting dementia. Their models overall accuracy ranges between 53.57% and 70.24%.

Joshi, et al. [16] tried different attribute-evaluation methods on the major risk factors of both Alzheimer's and Parkinson's diseases, which included both modifiable and non-modifiable risk factors. They used a relatively small dataset of fewer than 500 subjects from the ADRC and ISTAART studies [16]. Their attribute-evaluation methods included Chi-Squared, Gain Ratio, Info Gain, Relief F, and Symmetrical Uncertainty. They then applied several machine learning models, including Decision Tree, Random Forest (RF), and MLP to predict the patient's future status based on the defined risk factors. Their models did not detect dementia itself but instead classify subjects' diagnoses from three neurodegenerative diseases, which are AD, VaD, and Parkinson's.

Conversely, other studies aimed to predict dementia from neuroimaging data and in particular magnetic resonance imaging (MRI) or positron emission tomography (PET) scans of the brain. Ding, et al. [17] were able to predict Alzheimer's disease around six years before its diagnosis using fluorine 18 fluorodeoxyglucose PET images of the brain. They achieved 82% specificity at 100% sensitivity using a deep learning algorithm. In another study, Casanova, et al. [12] used both MRI images and cognitive tests to detect Alzheimer's risk using regularized logistic regression.

Although prediction using MRI or PET scans or even genetics data can be very accurate, it is not practical in many countries to scale such an approach for population screening, and it does not present direct links with potentially modifiable factors that could be taken into account by an individual patient to delay dementia.

There have been no published studies to date investigating machine learning approaches with larger datasets to link modifiable risk factors to dementia and therefore providing suggestions for treatment and lifestyle change based on multiple population-based longitudinal studies. Modern machine learning methods over and above those used in the aforementioned studies focusing on modifiable risk factors and larger datasets should be explored to determine if they can produce better predictions and Insight.

Moreover, using possibly interpretable models in clinical research is essential for intervention development and for gaining an understanding of the relationships and interactions between symptoms or risk factors and diagnosis. Interpretability is difficult to achieve using black-box models such as neural networks, which contains hidden layers, although they might yield higher prediction accuracy. The easiest way to achieve interpretability is through interpretable models such as linear and logistic regressions, decision trees, and Naïve Bayes [18]. Consequently, this paper focuses on such methods with modifiable risk factors as input variables trained and tested on datasets significantly larger than those reported upon to date.

C. The Alzheimer's Disease Neuroimaging Initiative (ADNI) Study

Early prediction of dementia requires tracking changes in cognitive ability over time. The ideal study type which can support this tracking is one yielding longitudinal data points. In longitudinal studies, data are collected on one or more variables repeatedly, over time, in contrast with cross-sectional studies, in which data are collected on one or more variables at a single time point [19] [20].

The Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu) is a longitudinal study that was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. Its primary goal has been to test whether MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) [21].

The dataset consists of three longitudinal studies on around 1900 participants in total. ADNI enrolls participants who are between the age of 55 and 90 and are either normal healthy older adults used as controls (CN), people with either early or late MCI, and people with AD. The cognitive-state diagnoses (as well as dementia status) rating assessment of the participants was also provided.

The ADNI data set has been widely used in many research studies [11], [12], [13]. However, none of the published research that has used the dataset to date has attempted a machine learning approach to predict dementia based on established modifiable risk factors or even to explore the dataset for other possible dementia risk factors. Most of the research has instead focused on using MRI and PET scans or genetic data to predict Alzheimer's disease.

Most previous studies using the ADNI dataset and other longitudinal studies in the dementia field have used a complete case analysis (CCA) [22], and thus they considered only the cases with complete data and removed the missing values [11] [15] [23] [24]. Moreover, as per [25], if there is an overall worsening trend in health over time, missing data can be imputed from the same subject using their other available data.

While researches have shown the importance of preventing or delaying dementia, which might be achieved by targeting known modifiable risk factors, few studies have applied machine learning approaches to selecting dementia's risk factors and predicting dementia status. However, some studies combined both modifiable and non-modifiable risk factors.

More research and work in this area would improve the early prediction of dementia and recommend actions that would possibly prevent or at least delay its onset by targeting only the non-modifiable risk factors. Using an interpretable machine learning approach on the attribute selection and prediction would help to predict dementia based on its modifiable risk factors.

The main research contribution of this study is a demonstration of the utility of interpretable machine learning methods for the purposes of predicting future cognitive status for an individual based on modifiable risk factors that have been already defined by the *Lancet* commission and the *Libra* index.

III. Methodology

This research follows one of the most widely used process models for predictive data analytics, which is the Cross-Industry Standard Process for Data Mining (CRISP-DM) model adapted from [26] (see Fig. 1). The project lifecycle phases, as illustrated in the diagram, are business understanding, data understanding, data preparation, modeling, evaluation, deployment, and monitoring. All phases are going to be included in this project except deployment and monitoring, which are beyond the scope of this research. Domain understanding has already been established in the background (Section II).

A. Understanding the ADNI Dataset

The ADNI dataset is extensive, containing hundreds of tables with different categories from primary patients' demographics to highly complicated genes and imaging datasets; however, not all tables were useful for the scope of this research. Therefore, an initial investigation of the dataset and its categories, subcategories, tables, fields, and their descriptions were needed. Fortunately, ADNI provides a data dictionary and an inventory that describe each table and its fields. The risk factors features are the independent variables while the diagnoses of the cognitive state are the independent variable, which might be one of three: cognitive normal (CN), mild cognitive impairment (MCI), and Dementia.

1) The modifiable risk factor attributes in ADNI: As the ADNI study dataset is extensive and consists of hundreds of tables and features under multiple categories, which may not be needed or useful for the aim of this research, the data dictionary, and the inventory were used to track only the necessary tables and features within them.

After reviewing the tables listed, the attributes related to dementia risk factors and diagnoses were selected. These attributes are listed in Table II. Attributes were selected from all ADNI cohorts except ADNI3 as the protocol of taking the medical history was different, and thus, not all features were available. A total of 1812 subjects were considered in the analysis.

2) Cross-Sectional vs. longitudinal data: As the dataset used in this research is longitudinal, another step was needed to understand the data through the study timeline. First, an understanding of how the data appear as cross-sectional, either at the baseline or at any single time point, was obtained. Then, a complete longitudinal view of the dataset was analyzed, including the differences between the main study parts (i.e., ADNI 1, Go, 2, and 3) and each visit's collected data.

B. *Data Preparation*

In this phase, the data were prepared for modeling by applying various data mining techniques to clean and to preprocess the data. This includes handling missing values, feature extractions, features transformation, and other tasks. Dealing with longitudinal data adds a complexity level to the preparation process because there could be various reasons and explanations for the data over time. A summary of the data preparation steps is shown in Fig. 2.

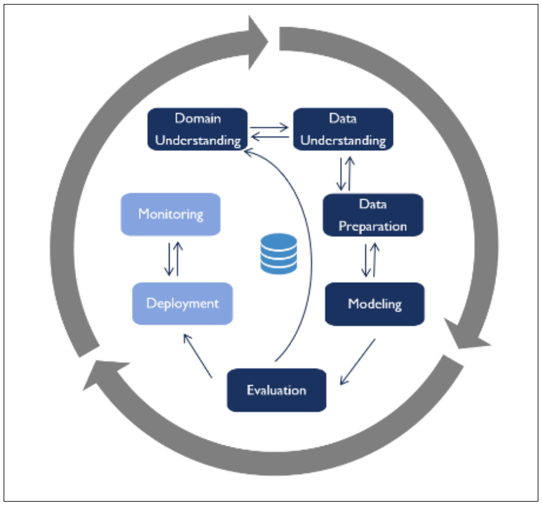


Fig. 1. CRISP-DM Model for the Project Phases (Adapted from [26]).

TABLE. II. MODIFIABLE RISK FACTORS AND DIAGNOSIS ATTRIBUTES

#	Risk Factor	Attributes Availability
Potentially Modifiable		
1	Low Education	Years of education
2	Hypertension	Detailed data available
3	Obesity	Can be calculated from weight\height
4	Smoking	Detailed data available
5	Depression	Detailed information
6	Diabetes	Check medical history and laboratory test results
7	Physical inactivity	Search relevant questioner's answers
8	Hearing loss	Search related terms on reported medical history
9	Social isolation	Search questioner's answers and related features (marital status, work)
10	Cognitive inactivity	Search relevant questioner's answers and related features
11	Chronic Heart disease	Check medical history
12	Alcohol use	Available
13	Chronic Kidney disease	Check medical history
Diagnosis		
14	Cognitive State	CN, MCI, and Dementia (available: baseline, follow up)

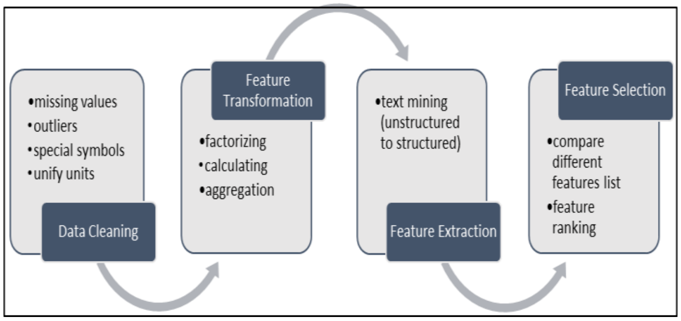


Fig. 2. A Summary of the Data Preparation Steps used in this Study.

1) Dealing with missing values in longitudinal data: Based on the ADNI study description, missing data were coded with -1 or -4. Typically, -4 is used for not applicable (i.e., data is not collected at a specific visit), and -1 is used for confirmed missing data. The detailed study schedule shows the data collected at each visit for each cohort group (i.e., CN, MCI, and AD).

To check the reason for missing data and to determine whether the data were missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR) [22] [25], the visit schedule descriptions, the visit registry table, and the exclusion tables were checked. The exclusion tables helped determine the reason for dropout, which might not be related to dementia, such as study partner availability, moving to another city, or not being willing to undergo MRI scans. For the available records missing data, several reasons were identified, and different actions were applied.

a) Missing Data Due to Scheduled Visit Design: In some cases, the data were missing because they were not collected during a visit (e.g., some visits were only for MRI imaging session; some data were collected only at the baseline). The missing data in these cases were considered MCAR and were imputed using the same patient's previous data, following the last observation carried forward (LOCF) [22] method.

- *Missing Height:* In the detailed ADNI visit schedules, participant's heights were only taken once during a screening visit, unlike their weights, which were repeatedly taken at each visit. Therefore, missing height data for each visit were filled in using a participant's screening visit height.

- *Missing Demographics and Medical History:* These data were collected only during the screening visit (repeated at the screening visit for each cohort, i.e., if a participant was included in ADNI1, 2, 3, medical history was taken at the screening visit for each cohort). The missing data were filled in using the same data for all visits (not imputation rather than fixed, although it might change, this is not recorded).

b) Missing Data Due to ADNI Study Stage Design: If the data were not collected during a specific ADNI stage, this means that the data were missing for all patients enrolled only during this stage. Therefore, only available or complete cases were considered. Examples of this include detailed smoking history, alcohol use, and medical history, which are not available for the ADNI3 study design. This led to selecting only ADNI1, ADNIGo, and ADNI2 cohorts for the study. Also, cognitive activity data were collected only beginning from ADNIGo, and thus, participants who were enrolled only in ADNI1 were excluded.

2) Feature transformation: Not all features have the desired format. Some new features must be calculated from existing ones, and some binary or categorical features must be factorized or encoded. Moreover, some features must be aggregated because they are repeated in multiple rows and could be defined as unique new features. The applied feature transformation included:

a) Unit Modifications: Height and weight units were not unified for all entries. Some were recorded as kg\cm, lbs\inch, lbs\cm, or kg\inch. All measurements were modified to the metric unit kg\cm.

b) Calculation: Some features needed to be calculated from other existing features. This may cause multicollinearity, which was reduced by selecting the best representative features which yield to better models results [11]. The calculated features were as follow:

- *BMI and Obesity:* Body mass index (BMI) was calculated based on the height and weight of the participants. Moreover, obesity was recorded when BMI >30 [27].

- *Social Isolation:* Social isolation level has been detected by calculating the available relative features, which are the marital status and retirement (as per [28], [29]).

- *Physical Activity:* Physical activity level has been calculated by adding up the related functional and physical assessment questioners' answers such as going shopping, playing games, and going out of the neighborhood.

c) Factorization: Visit codes were in a string format and were factorized to be numerical for simple computations and comparisons.

d) Aggregation: Structured medical description row-based fields were converted to binary column-based features (i.e., row for each condition per participant converted to 1 row with all conditions per participant).

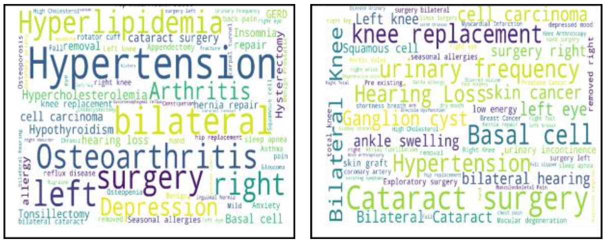
e) Normalization: For modeling purpose, numerical data has been normalized to range from 0 to 1 using the MinMaxScaler.

f) Encoding Categorical Features: Categorical features were encoded using dummy variables by converting the feature of k-categories to k-1 different dummy variables [30]. This was applied to the marital status and gender variables.

3) Feature extraction: Most risk factors available within the medical history description were text entries. These descriptions were entered as a free, unstructured text field with multiple variations of the same condition, which required some preprocessing to extract the features.

Some basic text mining techniques were applied to extract the previously defined risk factors and then to check for other possible factors. Using the NLTK package, stop words were removed, the text was converted to lower case, the most common terms and n-grams were selected, word clouds were plotted, and the known risk factor terms were searched and selected. Fig. 3 illustrates how medical history descriptions differ between those with dementia and others.

After applying text mining, each unstructured medical history text field was converted to a structured field (categorized), which is illustrated by Fig. 4.



(a) All Diagnosis (b) Only with Dementia

Fig. 3. Word Cloud of Most Common Medical History Descriptions

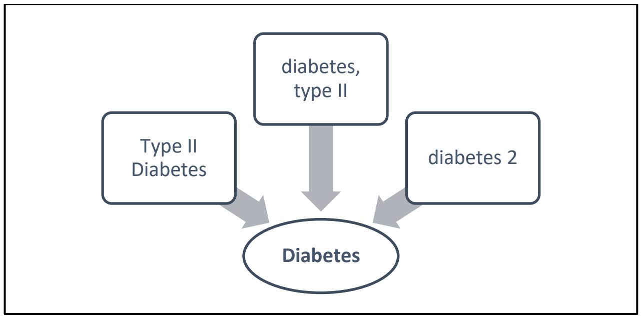


Fig. 4. Unstructured to Structured Medical Descriptions.

4) Feature selection: The previously defined features that were clinically approved to be relevant were selected. Both *Lancet* commission and *Libra* index modifiable risk factors features were considered in order to check, which gives better results.

5) Data integration: All selected tables were integrated and merged into a single table with all considered features.

C. Modeling

This research focused on interpretable modeling because of its importance for informing clinicians managing patients. Interpretable machine-learning classification models, such as the Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest. Both binary (dementia vs. non-dementia), and multi-class (CN vs. MCI vs. dementia) classifications were applied using the models.

1) Logistic regression: Logistic regression (LR) is an extension of linear regression and is used to solve classification problems. Basically, it is designed to solve binary classification problems where there are only two outcomes, but eventually, it is extended to support multi-classification, which is referred to as a multinomial logistic regression [26] [18]. A well-known method used to achieve a multinomial classification is using a set of one-versus-all models. For example, if there are *n* targets levels, *n* numbers of one-versus-all logistic regression models are created, and each model distinguishes between the features of one target level and all the others [26] [30].

2) Naïve bayes: In machine learning, the Naïve Bayes (NB) method serves as a probabilistic classifier that uses the Bayes' theorem of conditional probabilities [18] [26]. It assumes a strong (naïve) independence between features and calculates the class probabilities for each feature independently. The conditional probability of a class is the normalized class probability times the probability of each feature given by a class [18].

3) Decision tree: A decision tree (DT) is a tree-based model that splits the data repeatedly according to specific cutoff values in the features [18] [26]. Different subsets are created through splitting, separating instances to belong to one subset. The intermediate subsets are the internal nodes, while the final subsets are the leaf nodes. Decision trees are most useful if the relationship between the features and the target are nonlinear or if there are interactions between features.

4) Random forest (ensemble learning): The random forest (RF) model is an ensemble learning model that combines bagging, subspace sampling, and decision trees to create a more powerful model [26] [30]. The random forest model overcomes the overfitting problem of a decision tree, which is why it usually performs better.

A random forest model is a collection of decision trees in which each tree is slightly different from another. Once each individual decision tree model has been created (bagging), the ensemble makes predictions by returning the majority vote of the classifiers. This reduces the overfitting amount by averaging the results while maintaining the predictive power of each tree [30].

D. Evaluation

After the models are developed, the results were evaluated using multiple metrics and techniques to identify possible problems with overfitting and parameter tuning issues.

1) Confusion Matrix-Based performance measures: A confusion matrix is a convenient method used to comprehensively describe the performance of classification evaluations, which can be either binary or multi-class [26] [30] [31]. Most other metrics are derived from the basic components of the confusion matrix, which are the *True Positive (TP), True Negative (TN), False Positive (FP),* and *False Negative (FN),* and their percentage conversions. From these components, main evaluation measures such as accuracy, precision, recall, and F-score were calculated [31]. In this study, recall (sensitivity) is defined as the proportion of subjects who have dementia that are correctly classified. Precision is defined as the proportion of subjects who did not have dementia that are correctly classified. Accuracy is defined as the proportion of all subjects that are correctly classified, while F1 is the weighted average of precision and recall.

2) Sensitivity, specificity, and AUROC: The receiver operating characteristic (ROC) evaluates a model's true performance while considering all possible probability cutoffs (thresholds). The default threshold is 0.5; however, it could range from 0 to 1, and the classification results may change accordingly. The area under the ROC (AUROC) summarizes thresholds changes of both TPR (sensitivity) and FPR (1-specificity). The perfect fit is 1, the worst is 0, and the random prediction is 0.5.

IV. Experiments and Results

The experiments and analysis conducted for this research were applied using the following environments, tools, and libraries:

1) Environments Used: Python (3.6.4) and R.

2) Tools Used: Jupyter Notebook version 5.4.0, Google Colab (for faster modeling), and SPSS version 24 (for missing data mechanism and quickly find and explore).

3) Main Libraries and Packages: scikit-learn (for machine learning), NLTK (for text mining).

A. Model Validation

A balanced train-and-test split was applied using the *StratifiedKFold* to split the data once into a 75% training set and a 25% testing set. This ensures the same percentage of each class per group. Moreover, longitudinal data grouping by the participant was performed using the *GroupKFold***,** which is a special variant of cross-validation that takes into account the repeated measurements from the same subject and considers them as grouped data. Parameter tuning was achieved using nested cross-validation by applying *GridSearchCV* parameter tuning (inner loop) to cross-validation (outer loop).

B. Feature Selection

Using only statistically significant features either from the univariate analysis or per models was not sufficient because it decreased the accuracy from an average of 70% to 50%. This could be explained because there may be interactions between the features. The best feature selection was obtained using both the *Lancet* and the *Libra* index features. Using *Lancet* features alone gives an average of 59% accuracy only.

On the other hand, using *Libra* features only gives an average of 68% accuracy meaning that it is more comprehensive and predictive to the machine learning models, although combining it with *Lancet's* gives better results.

C. Cross-Sectional vs. Longitudinal Data Evaluation

Longitudinal data perform better than cross-sectional (baseline) data. However, the latest visit data gives the best results among them all.

Table III summarizes the results of the multi-class classifications on different data subsets for all models using ten-fold cross-validation.

TABLE. III. EVALUATION RESULTS SUMMARY (MULTI-CLASS)

Model	**LR**	**NB**	**DT**	**RF**
Longitudinal				
Accuracy	68.13%	57.99%	66.75%	**68.55%**
Precision	68.49%	65.38%	67.05%	**68.92%**
Sensitivity (Recall)	68.13%	57.99%	66.75%	**68.55%**
F1	68.18%	53.82%	66.75%	**68.61%**
Cross-Sectional (Baseline)				
Accuracy	**63.71%**	54.01%	60.34%	63.29%
Precision	64.39%	57.52%	60.30%	**65.60%**
Sensitivity (Recall)	**63.71%**	54.01%	60.34%	63.29%
F1	**63.11%**	47.39%	60.29%	61.92%
Latest Visit				
Accuracy	**77.00%**	66.77%	70.29%	71.57%
Precision	**76.76%**	66.13%	70.18%	70.73%
Sensitivity (Recall)	**77.00%**	66.77%	70.29%	71.57%
F1	**76.35%**	66.38%	70.15%	70.51%

As shown in the table, for longitudinal data, the best performance results are obtained by RF and reached around 68%. Moreover, using the baseline data alone, best results reached around 63-65% only.

Furthermore, using the latest visit data, the best performance results are obtained using LR and reached around 77%, which is the best among all data subsets. The overall differences between metrics are relatively small for all models.

Below figures illustrate the models' results of each longitudinal, baseline, and latest visit subsets respectively. The NB gives the least performance results for all data subsets.

For both longitudinal and baseline data, the LR and RF results are very similar for all metrics, followed by the DT, while the NB results are very lower, as shown in Fig. 5 and Fig. 6.

Fig. 7 shows the latest visit results, where it is clear how the LR outperformed the other models for all metrics. The DT and RF results for this subset are relatively similar to each other.

As shown, considering only the latest visit subset gives better evaluation results for all models, followed by the longitudinal, and finally, the baseline subset. This performance difference is clearly illustrated in Fig. 8 for the LR model.

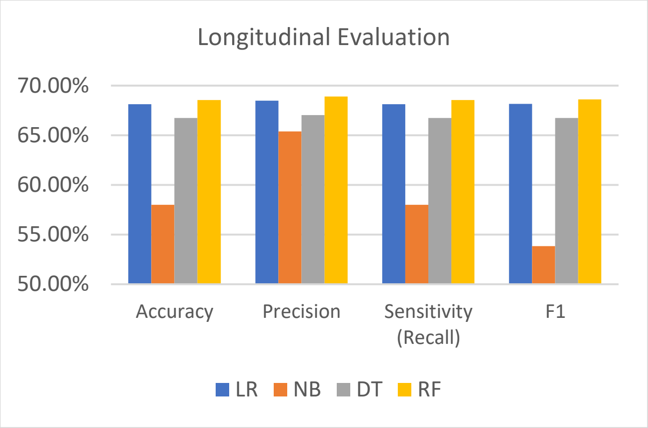


Fig. 5. Longitudinal Evaluation - (Multi-Class).

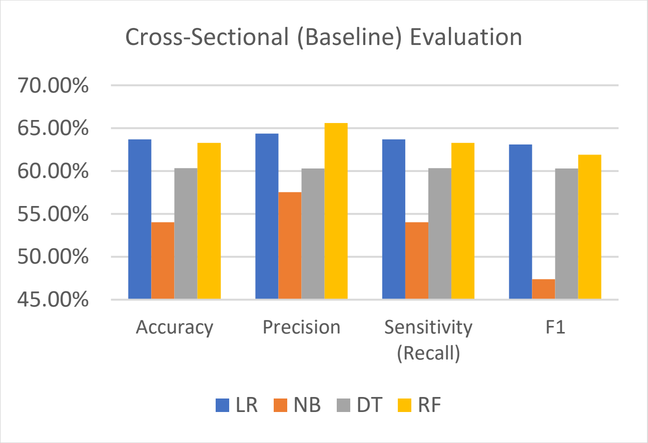


Fig. 6. Cross-Sectional (Baseline) Evaluation - (Multi-Class).

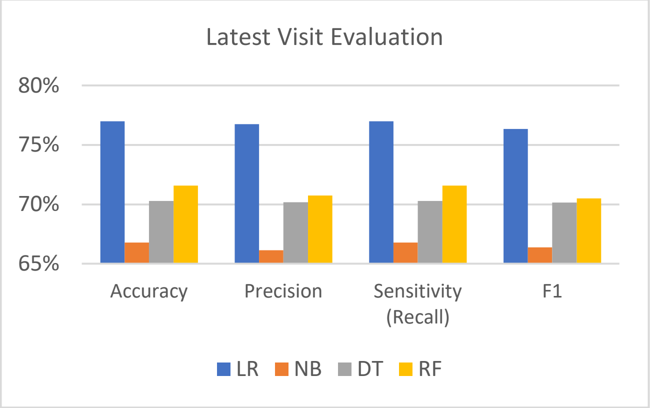


Fig. 7. Latest Visit Evaluation-(Multi-Class).

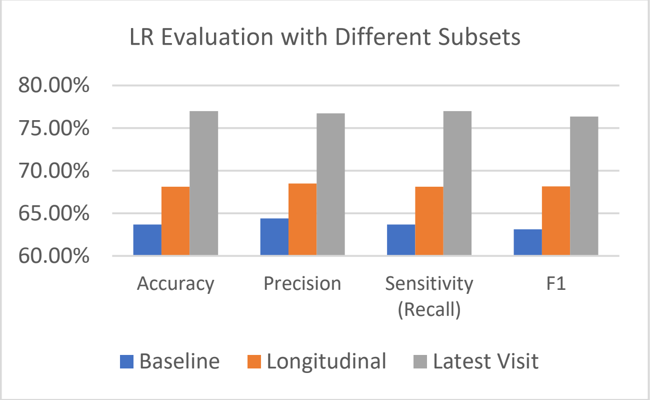


Fig. 8. LR Evaluation Comparison of different Subsets-(Multi-Class).

The area under the ROC curve (AUROC) of the dementia class was 96% for both top models (LR and RF) as illustrated in Fig. 9 and Fig. 10, respectively. Although the AUROC of MCI and CN classes are lower than the dementia class (CN: 86%-88% and MCI: 67%-78% for LR and RF, respectively), the AUROC of the dementia is the more important in this study.

D. Binary vs. Multi-Class Evaluation

In addition to the multi-class classification (CN vs. MCI vs. dementia), binary classification (dementia vs. non-dementia) has been applied using all previous models. Binary classifications outperformed multi-class classification, although they are less informative. Table IV and Fig. 11 demonstrates the results for both the binary and the multi-class classifications for the two top performed models (i.e., LR and RF). While the multi-class models' result reaches 70% only, the binary models reach around 92% on all metrics. The LR results are better than the RF for all metrics.

E. Overfitting Check and Model Generalization

All models have been checked against overfitting by comparing the training and testing accuracies. The difference between the train and test accuracies ranged between 0 (LR and NB), 0.02 (RF), and 0.04 (DT), which is considered small and acceptable. Moreover, to ensure cross-validation generalization, the standard deviation of the accuracy for all folds has been checked. This ranged between 0.02 and maximum 0.03, which is all considered small and acceptable. Table V shows the detailed results of the models check.

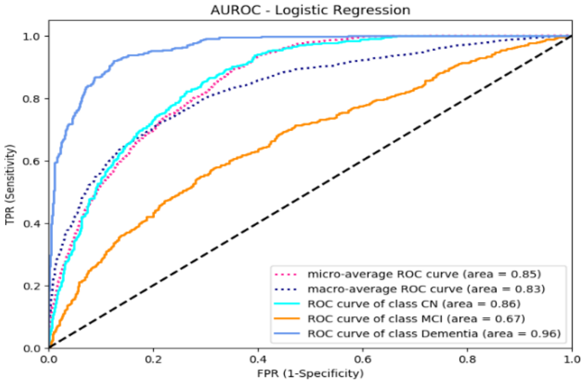


Fig. 9. AUROC for LR–Multi-Class Classification (Longitudinal).

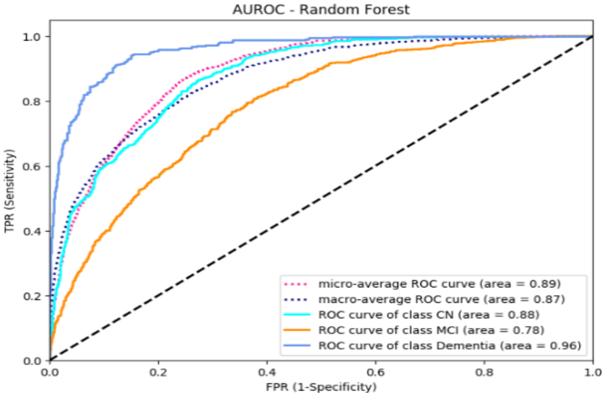


Fig. 10. AUROC for RF–Multi-Class Classification (Longitudinal).

TABLE. IV. EVALUATION RESULTS OF BINARY VS. MULTI-CLASS

Model	LR	RF	LR	RF
	Binary		*Multi-class*	
Accuracy	**91.53%**	91.24%	77.00%	71.57%
Precision	**91.34%**	90.95%	76.76%	70.73%
Sensitivity	**91.53%**	91.24%	77.00%	71.57%
F1	**91.41%**	91.01%	76.35%	70.51%

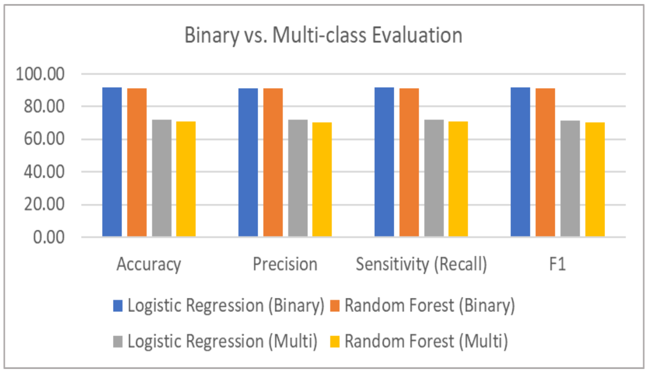


Fig. 11. Binary vs. Multi-Class Evaluation.

TABLE. V. OVERFITTING AND MODELS GENERALIZATION CHECK

Results\ Model	**LR**	**NB**	**DT**	**RF**
Train-Test Split				
Training Accuracy	0.92	0.87	0.94	0.93
Testing Accuracy	0.92	0.87	0.90	0.91
Training - Testing Accuracy	**0.00**	**0.00**	**0.04**	**0.02**
Cross-Validation				
CV Folds Accuracy Mean	0.91	0.87	0.88	0.90
Standard Deviation	**0.02**	**0.03**	**0.02**	**0.03**

F. *Feature Importance*

Feature importance was calculated using different models. For linear models, it was calculated based on the absolute values of the coefficients. For tree-based models, it was calculated based on the model's feature importance. Table VI summarizes the risk factors' importance for each model. From the table, it is clearly shown that BMI, Cognitive Activity, and Physical Activity feature importance are top across most models. The feature importance was extracted from the best performing models which combine both *Lancet* and *Libra* lists.

G. *Evaluation Summary*

The best classification results were obtained using both the *Lancet* and the *Libra* risk factor lists, considering the longitudinal data set which outperformed the cross-sectional baseline one. Moreover, using data of the most recent visits only provided even better results than using the whole longitudinal set.

In some cases, it is important to detect whether a person has dementia or not, while in other cases, the exact cognitive state is needed. Therefore, both binary and multi-class classifications have been applied. The binary classification yielded to about 92% accuracy, while the multi-class classification yielded to a 77% accuracy using logistic regression, followed by random forest with 92% and 70%, respectively. The area under the ROC of the dementia class was nearly perfect at 96% for both models.

TABLE. VI. FEATURE IMPORTANCE FOR TOP MODELS

#	**Feature\ Model**	**Feature Importance Order**				
		LR	**NB**	**DT**	**RF**	**All Models**
1	BMI	3	7	3	3	16
2	Cognitive Activity	1	13	2	1	17
3	Physical Activity	2	12	1	2	17
4	Smoking	12	1	4	5	22
5	Alcohol	4	2	9	11	26
6	Heart	5	4	12	8	29
7	Kidney	6	3	7	13	29
8	Depression	8	9	8	6	31
9	Hearing Loss	9	6	6	12	33
10	Education	11	14	5	4	34
11	Hypertension	7	10	10	10	37
12	Diabetes	13	5	14	9	41
13	Social Isolation	14	11	11	7	43
14	Cholesterol	10	8	13	14	45

Although features importance was not identical for all models, the top three features importance were identical for the two top performed models (i.e., LR and RF), which is a sign of model's stability. Furthermore, as this is an observational study analysis, the feature importance of each model does not claim any causality of dementia or MCI. The importance derived from the available data may not be representative of a wider population.

The best obtained results of this study were either competitive or even better than the results obtained by other previous studies which used MRI data and machine learning or deep learning methods [11], [12], [13]. Their best overall accuracies range between 65% and 92%. Moreover, the results of this study were better than a previous study that used a machine learning approach with modifiable risk factors, where their best accuracy reaches 75.24% only [15]. However, this is not considered as a complete comparison as the other studies used different datasets.

V. CONCLUSION

A. *Achievements of the Research Objectives*

The research discussed and evaluated in the previous sections aims to use different interpretable machine-learning classification models to detect dementia based on its modifiable risk factors only. It explored and applied *Lancet*, and *Libra* lists of modifiable risk factors on the ADNI dataset, which is as far as known have not been applied on this dataset using machine learning approaches.

The best classification results were obtained using both the *Lancet* and the *Libra* risk factor lists. Considering the longitudinal data set outperformed the cross-sectional baseline one. Moreover, using data of the most recent visits only provided even better results than using the whole longitudinal set.

The binary classification yielded to about 92% accuracy, while the multi-class classification yielded to a 77% accuracy using logistic regression, followed by random forest with 92% and 70%, respectively. Furthermore, the best achieved overall accuracies were either competitive to or better than previous studies results.

B. *Limitations*

This research involved an experimental analysis of an observational study based on the ADNI dataset, and there is no claim to present causations. The ADNI study was not primarily designed to address the modifiable risk factors; thus, it may lack some useful features, especially during the early and middle life courses. Social isolation and physical activities are not explicitly addressed by the study, and the results may be more accurate if more detailed data for these factors were collected. Medical history and other important useful demographic features, such as occupation, were collected as free text and were not categorized in a structured format during the data collection stage, which may have helped make the analysis simpler and more accurate.

ACKNOWLEDGMENT

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List .pdf.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

REFERENCES

[1] G. Livingston, A. Sommerlad, V. Orgeta, S. G. Costafreda, J. Huntley, D. Ames, et al., "Dementia prevention, intervention, and care," The Lancet, vol. 390, pp. 2673-2734, 2017.

[2] A. P. Association, Diagnostic and statistical manual of mental disorders (DSM-5®): American Psychiatric Pub, 2013.

[3] S. J. Vos, M. P. Van Boxtel, O. J. Schiepers, K. Deckers, M. De Vugt, I. Carriere, et al., "Modifiable risk factors for prevention of dementia in midlife, late life and the oldest-old: validation of the LIBRA Index," Journal of Alzheimer's Disease, vol. 58, pp. 537-547, 2017.

[4] K. Yaffe, "Modifiable risk factors and prevention of dementia: What is the latest evidence?," JAMA Internal Medicine, vol. 178, pp. 281-282, 2018.

[5] K. Deckers, M. P. van Boxtel, O. J. Schiepers, M. de Vugt, J. L. Muñoz Sánchez, K. J. Anstey, et al., "Target risk factors for dementia prevention: a systematic review and Delphi consensus study on the evidence from observational studies," International journal of geriatric psychiatry, vol. 30, pp. 234-246, 2015.

[6] O. J. Schiepers, S. Köhler, K. Deckers, K. Irving, C. A. O'donnell, M. van den Akker, et al., "Lifestyle for Brain Health (LIBRA): a new model for dementia prevention," International journal of geriatric psychiatry, vol. 33, pp. 167-175, 2018.

[7] C. A. O'Donnell, V. Manera, S. Köhler, and K. Irving, "Promoting modifiable risk factors for dementia: is there a role for general practice?," The British journal of general practice : the journal of the Royal College of General Practitioners, vol. 65, pp. 567-568, 2015.

[8] E. National Academies of Sciences and Medicine, Preventing cognitive decline and dementia: A way forward: National Academies Press, 2017.

[9] S. Tariq and P. A. Barber, "Dementia risk and prevention by targeting modifiable vascular risk factors," Journal of Neurochemistry, vol. 144, pp. 565-581, 2018/03/01 2018.

[10] X.-H. Hou, L. Feng, C. Zhang, X.-P. Cao, L. Tan, and J.-T. Yu, "Models for predicting risk of dementia: a systematic review," Journal of Neurology, Neurosurgery, and Psychiatry, 2018.

[11] B. Bratić, V. Kurbalija, M. Ivanović, I. Oder, and Z. Bosnić, "Machine Learning for Predicting Cognitive Diseases: Methods, Data Sources and Risk Factors," Journal of Medical Systems, vol. 42, p. 243, 2018/10/27 2018.

[12] R. Casanova, F.-C. Hsu, K. M. Sink, S. R. Rapp, J. D. Williamson, S. M. Resnick, et al., "Alzheimer's Disease Risk Assessment Using Large-Scale Machine Learning Methods," PLOS ONE, vol. 8, p. e77949, 2013.

[13] R. Casanova, R. T. Barnard, S. A. Gaussoin, S. Saldana, K. M. Hayden, J. E. Manson, et al., "Using high-dimensional machine learning methods to estimate an anatomical risk factor for Alzheimer's disease across imaging databases," NeuroImage, vol. 183, pp. 401-411, 2018/12/01/ 2018.

[14] M. Baumgart, H. M. Snyder, M. C. Carrillo, S. Fazio, H. Kim, and H. Johns, "Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective," Alzheimer's & Dementia, vol. 11, pp. 718-726, 2015/06/01/ 2015.

[15] J. O'Donoghue, M. Roantree, and A. McCarren, "Variable interactions in risk factors for dementia," in 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS), 2016, pp. 1-10.

[16] S. Joshi, P. D. Shenoy, K. Venugopal, and L. Patnaik, "Classification of neurodegenerative disorders based on major risk factors employing machine learning techniques," International Journal of Engineering and Technology, vol. 2, p. 350, 2010.

[17] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, et al., "A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain," Radiology, p. 180958, 2018.

[18] C. Molnar, Interpretable machine learning: A guide for making black box models explainable: Leanpub, 2019.

[19] P. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger, Analysis of Longitudinal Data vol. 25: OUP Oxford, 2013.

[20] S. Menard, Handbook of longitudinal research: Design, measurement, and analysis: Elsevier, 2007.

[21] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 27, pp. 685-691, 2008.

[22] M. Nakai and W. Ke, "Review of the methods for handling missing data in longitudinal data analysis," International Journal of Mathematical Analysis, vol. 5, pp. 1-13, 2011.

[23] F. Ben Bouallègue, D. Mariano-Goulart, P. Payoux, and I. Alzheimer's Disease Neuroimaging, "Comparison of CSF markers and semi-quantitative amyloid PET in Alzheimer's disease diagnosis and in

cognitive impairment prognosis using the ADNI-2 database," Alzheimer's research & therapy, vol. 9, pp. 32-32, 2017.

[24] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, and A. s. D. N. Initiative, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," Neuroimage, vol. 55, pp. 856-867, 2011.

[25] J. M. Engels and P. Diehr, "Imputation of missing longitudinal data: a comparison of methods," Journal of Clinical Epidemiology, vol. 56, pp. 968-976, 2003/10/01/ 2003.

[26] J. D. Kelleher, B. Mac Namee, and A. D'arcy, Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies: MIT Press, 2015.

[27] G. Piumatti, S. C. Moore, D. M. Berridge, C. Sarkar, and J. Gallacher, "The relationship between alcohol use and long-term cognitive decline in middle and late life: a longitudinal analysis using UK Biobank," Journal of Public Health, vol. 40, pp. 304-311, 2018.

[28] E. Y. Cornwell and L. J. Waite, "Measuring social isolation among older adults using multiple indicators from the NSHAP study," The journals of gerontology. Series B, Psychological sciences and social sciences, vol. 64 Suppl 1, pp. i38-i46, 2009.

[29] J. Holt-Lunstad, T. B. Smith, M. Baker, T. Harris, and D. Stephenson, "Loneliness and Social Isolation as Risk Factors for Mortality: A Meta-Analytic Review," Perspectives on Psychological Science, vol. 10, pp. 227-237, 2015/03/01 2015.

[30] A. C. Muller and S. Guido, Introduction to machine learning with Python: a guide for data scientists: O'Reilly Media, 2017.

[31] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, pp. 427-437, 2009.

Most Valuable Player Algorithm for Solving Minimum Vertex Cover Problem

Hebatullah Khattab[1], Ahmad Sharieh[2], Basel A. Mahafzah[3]
Department of Computer Science
King Abdulla II School of Information and Technology
The University of Jordan, Jordan

***Abstract*—Minimum Vertex Cover Problem (MVCP) is a combinatorial optimization problem that is utilized to formulate multiple real-life applications. Owing to this fact, abundant research has been undertaken to discover valuable MVCP solutions. Most Valuable Player Algorithm (MVPA) is a recently developed metaheuristic algorithm that inspires its idea from team-based sports. In this paper, the MVPA_MVCP algorithm is introduced as an adaptation of the MVPA for the MVCP. The MVPA_MVCP algorithm is implemented using Java programming language and tested on a Microsoft Azure virtual machine. The performance of the MVPA_MVCP algorithm is evaluated analytically in terms of run time complexity. Its average-case run time complexity is ceased to $\boldsymbol{\Theta(I(|V|+|E|))}$, where *I* is the size of the initial population, |*V*| is the number of vertices and |*E*| is the number of edges of the tested graph. The MVPA_MVCP algorithm is evaluated experimentally in terms of the quality of gained solutions and the run time. The experimental results over 15 instances of DIMACS benchmark revealed that the MVPA_MVCP algorithm could, in the best case, get the best known optimal solution for seven data instances. Also, the experimental findings exposed that there is a direct relation between the number of edges of the graph under test and the run time.**

Keywords—Most valuable player algorithm; minimum vertex cover problem; metaheuristic algorithms; optimization problem

I. Introduction

In general, many heuristic and metaheuristic algorithms were used to solve many optimization problems; such as Minimum Vertex Cover Problem (MVCP), Traveling Salesman Problem (TSP), 15 puzzle problem, task scheduling, software testing, and non-optimization problems [1-4]. Examples of heuristic algorithms are A* heuristic search algorithm and iterative deepening A* (IDA*) heuristic search algorithm [5]. Examples of metaheuristic algorithms are sea lion optimization, humpback whale optimization, Genetic Algorithm (GA), Ant Colony Optimization (ACO), Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), and Chemical Reaction Optimization (CRO) [6-14].

The MVCP is a combinatorial optimization problem in computer science. It is a classic example of an NP-hard optimization problem. The MVCP is the problem of finding the smallest set of vertices such that at least one endpoint of each edge of the tested graph belongs to that set [15].

The Vertex Cover Problem (VCP) can be defined as follows: let *G* = (*V*, *E*) be an undirected graph with set *V* of vertices and set *E* of edges. If *S* is a subset of *V* ($S \subseteq V$) and (*x*, *y*) is an edge in *G*, then *S* is the cover of *G* if either $x \in S$ or $y \in S$ or both [15]. The MVCP is the problem of finding a subset *S* such that |*S*| is the minimum. The MVCP can be formulated as in (1).

$$min \sum_{x \in V} w_x \quad (1)$$

Subject to: $|E| = \sum_{(x,y) \in E} c_{(x,y)}$

where $w_x = \begin{cases} 1 & if\ x\ exists\ in\ the\ solution \\ 0 & if\ x\ does\ not\ exist\ in\ the\ solution \end{cases}$

and

$$c_{(x,y)} = \begin{cases} 1 & if\ at\ least\ x\ or\ y\ exists\ in\ the\ solution \\ 0 & if\ neither\ x\ nor\ y\ exists\ in\ the\ solution \end{cases}$$

For the sake of clarity, Fig. 1 depicts an illustrative example for the MVCP. In this figure, the graph *G* consists of 6 vertices and 6 edges. For instance, the subset {1, 2, 5, 6} represents a cover for *G*, but it is not the minimum. In fact, the subsets {1, 5} and {1, 6} are the minimum ones.

Many real-life problems and applications can be developed as MVCP. Therefore, many scientists have been encouraged to create higher attempts to discover efficient solutions. Networks and communications [16], engineering [17], and bioinformatics [18] are some of the real-world applications that they were represented and solved as MVCP.

To solve the MVCP, distinct kinds of algorithms were developed to introduce worthy solutions. Several exact [19, 20], heuristic [21-23], and metaheuristic [10-14, 24] algorithms were presented to achieve the purpose of solving the MVCP. The exact algorithms always find the optimal solution to the optimization problems if the problem size is comparatively small. This is because, in a feasible time, the optimal solution can be achieved. But once the problem size starts to increase, heuristic and metaheuristic algorithms are needed. Both of these types of algorithms can find a solution as close as possible to the optimal solution. Maximum Degree Greedy (MDG), Vertex Support Algorithm (VSA), and New Modified Vertex Support Algorithm (NMVSA) [21, 22] are some of the heuristic algorithms that were introduced specifically for the MVCP.

Many researchers adapted metaheuristic algorithms to handle MVCP. The GA was used to solve the MVCP in [10, 11]. The ACO algorithm also tackled the MVCP in [12, 13]. Furthermore, it was addressed by the CRO algorithm in [14].

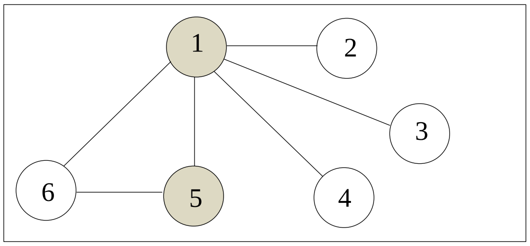


Fig. 1. A Graph *G* with 6 Vertices and 6 Edges, where MVC = {1, 5}.

The Most Valuable Player Algorithm (MVPA) is a recent metaheuristic algorithm that was introduced in 2017 for solving optimization problems [25]. The algorithm is inspired by sport, where the players make up teams, then they compete (in teams) together to win the championship. They also compete for the best player award on an individual basis.

In this paper, the MVPA is adapted and implemented for treating the MVCP. This algorithm is implemented using Java programming language and executed on a Microsoft Azure virtual machine. The performance of the implemented algorithm is evaluated analytically in terms of run time complexity and cost function metrics. Furthermore, it is evaluated experimentally in terms of solution quality and run time. The experiments have been conducted using different DIMACS benchmark instances for MVCP.

The structure of this paper is summarized as follows: Section II reviews some of the related work. In Section III, the MVPA is briefly lunched and explained. Section IV shows how the MVPA is tailored to the MVCP. Section V analytically evaluates the implemented MVPA. Section VI shows and discusses the experimental outcomes. Finally, Section VII concludes the conducted work and suggests some future work.

II. Related Work

As several essential applications are depicted as MVCP, it has been the heart of extensive research. In [26], a new local search algorithm that is named NuMVC (New Minimum Vertex Cover) has been introduced to tackle the MVCP. The primary concept for NuMVC was to confront some weaknesses, which normally occur in the local search algorithms, linked to the exchange of vertices and weight of edges. NuMVC has come up with new approaches to address these weaknesses.

VEWLS (Vertex Edge Weighting Local Search) algorithm is presented in [27]. This algorithm integrates the vertex weighting scheme with the edge weighting scheme. In comparison with the NuMVC algorithm, VEWLS performance was evaluated. The findings showed that the VEWLS algorithm was superior to the NuMVC algorithm.

Moreover, GA has been used to solve the MVCP in [10]. The primary objective of this research was to demonstrate the effects of growing the population size. The results exposed that the number of generations required getting the optimal solution decreases as the population size increases.

The ACO algorithm has been exploited in [13] to solve the minimum weight VCP. In this research, a heuristic strategy has been put forward to rule out suspicious elements to correct the pheromone. This strategy is based on extracting information related to the best solution. This information enhanced the canonical ACO algorithm by avoiding the premature trapping of the local optima. The findings indicated a noticeably shorter time, while the results achieved were somewhat better.

A version of CRO algorithm called Hybrid Chemical Reaction Optimization Algorithm (HCROA) was designed to solve the minimum vertex cover problem for an undirected graph in [14]. In this algorithm, a greedy approach is adopted for implementing the main reactions operators. The HCROA was compared with genetic algorithm and branch and bound approach. The comparison was in terms of the number of iterations that were executed to reach the optimal solution. The results exposed that HCROA outperformed the genetic algorithm and the branch and bound approach.

There are not many research work so far as the MVPA is concerned. This is due to the fact that MVPA is invented so recently [25]. However, the same author who introduced the MVPA has investigated his algorithm to tackle the optimal design of circular antenna arrays for maximum sidelobe levels reduction [28]. The results of testing the proposed algorithm showed that it is superior to many other counterpart algorithms.

In [29], a comparison has been conducted between the MVPA and other sport-inspired metaheuristic algorithms. All compared algorithms have been tested using unimodal and multimodal problems. The MVPA has been proved to be the best algorithm regarding unimodal problems. For the multimodal problems, it has been the best together with two other algorithms.

III. Most Valuable Player Algorithm

As previously stated, Bouchekara has recently introduced the MVPA in 2017 [25]. He inspired the idea of the MVPA from the team-based sports; where all participated players are grouped in teams. Algorithm 1 illustrates the MVPA phases.

The inputs of the MVPA are the *problem size* (the dimension of the tackled problem), *players size* (the number of players), *teams size* (the number of teams), and *MaxNFix* (the maximum number of fixtures (iterations)). The Most Valuable Player (MVP) that represents the best-obtained players is the output of the MVPA.

The MVPA begins with the initial phase. In this phase, the initial population of players is randomly created. In the main phase, all other phases are executed and repeated until the stop condition is satisfied. The first step that is executed in the main phase is the distribution of the players of the population into teams. The competition phase iterates for all teams. For each selected team, two types of competitions are carried out, *individual* and *team* competitions. In the individual competition, each player of the selected team tries to improve his sporting skills to be the best player in his team and the league. Concerning the team competition, it is performed among the competed teams. Each selected team plays against another randomly picked team. As a result of this play, the players of the selected team follow a certain mechanism to update their skills.

Algorithm 1: MVPA

```
Inputs: problem size, players size, teams size, and MaxNFix.
Output: MVP
Initial Phase:
    Creation of the initial population of players
Main Phase:
    Distribution of population players in teams
Do
    Competition Phase:
    for all teams
        Individuals Competition Phase
        Team Competition Phase
        Bound Checking Phase
    end for
    Greediness Phase
    Elitism Phase
    Duplicates Removing Phase
Until the stop criterion is satisfied
```

It must be mentioned that the player skills have lower and upper bounds, and in each competition, all players are constantly seeking to improve their skills. So, if the improvement tries updating the players' skills out of these bounds, these skills must be brought back to their bounds. This checking of the player's skills bounds is regularly performed in the bound checking phase.

At the end of the competition phase, the new population of players is shaped. This new population faces the greediness phase where the player who only got better skills is accepted, otherwise, his skills remain without accepting the conducted changes. In the elitism phase, a specific number of worst players in the new population are replaced with the same number of elite players who have the best skills. As a final phase, any duplicated player is replaced by a player from the winning teams.

As mentioned beforehand, the main phase iterates until the stop condition is satisfied. In [25], this condition is determined by the number of iterations which equals a specific number that is assigned to the *MaxNFix*. All information about the MVPA designing details can be found in [25].

IV. Most Valuable Player Algorithm for the Minimum Vertex Cover Problem

The first step of adapting the MVPA for the MVCP is to align its main concepts of the MVPA with the MVCP. Table I shows the main concepts of the MVPA and their related meanings of the MVCP. The *player* concept in the MVPA represents a solution of the graph considered. In order to implement a solution, let a graph $G = (V, E)$, then the vector $a = (a_1, a_2, ..., a_{|V|})$ where $a_i \in \{0, 1\}$, is a binary vector that represents the solution. If the i^{th} vertex contributes to covering the graph G, then $a_i = 1$, otherwise, $a_i = 0$. In consequence, the number of ones in the binary vector represents the solution size which is noted as *player skills* in the MVPA. To clarify the idea, consider the following example. For instance, for a graph G with 6 vertices, if a = (1, 1, 0, 1, 0, 1), then the solution size is 4 and vertices 1, 2, 4 and 6 cover all edges of the graph G.

The *problem size* in the MVPA corresponds to the number of vertices in the graph considered for the MVCP; which represents the length of the created binary vector.

In the MVCP, the number of initial solutions that are created to form the initial population corresponds to the *players size* concept in the MVPA. When the initial population's solutions are distributed over teams, the solution which has the minimum size in a team represents the *Franchise player*. However, the solution which has the minimum size in all teams is considered the *most valuable player*. The best-gained solution is the *most valuable player* after all the *MaxNFix* iterations.

The adaptation of MVPA for the MVCP is described in Algorithm 2. As inputs, the algorithm requires a graph to be tested, it is denoted as G. The initial population size which represents the number of initial solutions is symbolized by I. Variable T stands for the number of groups that the population solutions will be distributed into. The maximum number of times that the algorithm should iterate is denoted by M. The best solution obtained after the M iterations is represented by the Minimum Vertex Cover (*MVC*) which is the output of Algorithm 2.

In regards to MVPA phases, they are adapted for the MVCP firstly by assembling the competition, bound checking, greediness, and elitism phases in the main phase. Duplicates removing phase is ignored because of what will be explained after a while. Besides, the original order of the MVPA phases is modified to meet the needs of adaptation.

In the initial phase, as in line 1 of Algorithm 2 shows, I of initial solutions are randomly created to form the initial population. The initial solutions are created using a *Random Bit-Vector* (RBV) approach [30]. In RBV, the solution binary vector is made up by assigning each vertex value of 0 or 1 based on a generated random number. If this number is greater than a predefined constant, then the value of the vertex will be 1, or 0 otherwise. Calculating the sizes of these initial solutions is the next step. After then, the best solution (i.e. the solution with the minimum size) of the initial population is determined as illustrated in line 2.

TABLE. I. The MVPA Concepts for the MVCP

MVPA Concept	**MVCP Meaning**
Player	Solution.
Player skills	Solution size.
Problem size	Tested graph size (the number of its vertices).
Players size	The number of initial solutions (initial population size).
Franchise player	The solution with the minimum size in a team.
Most valuable player	The solution with the minimum size in all teams.

Algorithm 2: MVPA_MVCP

```
Inputs   Graph $G$ ($V$, $E$),
         $I$ (initial population size),
         $T$ (number of groups) and
         $M$ (maximum number of iterations)
Output   $MVC$ (minimum vertex cover; i.e. the best solution
         obtained)

     // Initial Phase
1    Create the initial population ($P$) by generating $I$ random
     initial solutions and calculate the sizes of these solutions.
2    Find the best solution in $P$ (i.e. the solution with
     minimum size) and denoted as $L$.
     // Main Phase
3    for $f$ = 1 to $M$
4        Distribute $P$ solutions over $T$ groups
         // Competition Phase
5        for $i$ = 1 to $T$
6            Retrieve group $i$ ($g_i$)
7            Pick randomly group $j$ ($g_j$), given that $i \neq j$
8            Find the best solutions in $g_i$ and $g_j$. Denote them
             as $B_i$ and $B_j$, respectively.
             // Individual Competition Phase
9            for each solution ($X$) in $g_i$ Do
10               for each vertex $d$ of $X$  Do
11                   $X_d = X_d + rand \times (B_{i-d} - X_d) + 2 \times rand \times (L_d - X_d)$
                     //Bound Checking Phase - Stage 1
12                   if $X_d \leq 0$ then $X_d = 0$, else $X_d = 1$
13               end for
                 // Greediness Phase - Stage 1
14                   if the new $X$ cover graph $G$ and its size
                       less than the original $X$ size then
15                       Accept the new $X$
16                   else
17                       keep the original $X$
18                   end if
19           end for
             // Team Competition Phase
20           Calculate the probability of winning $g_i$ against $g_j$
             and $g_j$ against $g_i$
21           for each solution ($X$) in $g_i$ Do
22               for each vertex $d$ of $X$  Do
23                   if $g_i$ wins against $g_j$
24                       $X_d = X_d + rand \times ( X_d - B_{i-d})$
25                   else
26                       $X_d = X_d + rand \times ( B_{i-d} - X_d)$
27                   end if
                     //Bound Checking Phase - Stage 2
28                   if $X_d \leq 0$ then $X_d = 0$, else $X_d = 1$
29               end for
                 // Greediness Phase - Stage 2
30               if the new $X$ cover graph $G$ and its size    less
                   than the original $X$ size then
31                   Accept the new $X$
32               else
33                   keep the original $X$
34               end if
35               end for
36           end for
             // Elitism Phase
37           Recollect the solutions from all groups in the
             population ($P$).
38           Sort $P$ solutions based on their sizes
39           Replace one-third of the worst solutions with one-
             third of the best solutions
40       end for
41       Output the best solution as $MVC$.
```

The main phase, which is represented in lines 3-40, iterates M times. The first step in each time is to subsequently spread the population solutions across the T groups as depicted in line 4. Shortly afterwards, the competition phase (lines 5-36) starts with retrieving the group that is due to be processed (g_i) as shown in line 6. In line 7, another group (g_j) is randomly retrieved to confront g_i in the team competition phase, where g_i and g_j should be different. Before delving in executing any of the competition phases (individual or team), it is needed first to find the best solution in g_i and g_j as exposed in line 8. Concerning the individual competition phase, it extends between lines 9 and 19. In this phase, each solution of g_i undergoes an improvement attempt to minimize its size using the equation in line 11 [25]. In this equation, the vertices values of each solution are updated based on the values of the vertices of the best solution in the g_i (B_i) and the best solution in population P (L). Regarding the team competition phase, its steps are allocated in lines 20-35. Its first step is to determine the winner group by calculating the probability of winning g_i against g_j and g_j against g_i. Equation (2) is used to calculate these probabilities [25].

$$Pr_{g_a Wins g_b} = 1 - \frac{(sizeN(g_a))^k}{(sizeN(g_a))^k + (sizeN(g_b))^k} \quad (2)$$

where g_a and g_b are any competed groups, k is a constant and $sizeN(g_a)$ is the normalized size of g_a's solutions sizes that is calculated as in (3) [25].

$$sizeN(g_a) = size(B_a) - \min(size(B_1), size(B_2), \dots, size(B_T)) \quad (3)$$

On the consequence of the winner group determination, the vertices values of g_i solution are updated using either the equation in line 24 or in line 26 [25].

As stated formerly, the solution vertices values are restricted to be 0 or 1. However, when the equations in lines 11, 24, and 26 are used to update the values of the vertices, some of the obtained values differ from 0 or 1. So, in the bound checking phase, these values must be checked and brought back to 0 or 1. Specifically, when the value acquired of applying any of these equations is less than 1, then the vertex value is determined to be 0. Otherwise, it is considered to be 1. Since the values of the vertices are updated in both individual and team competition phases, the bound checking phase is executed twice, once after each updating process. This is illustrated in lines 12 and 28.

Intuitively, after each updating, the sizes of the updated solutions are re-calculated. Accepting these solutions essentially is based on the fact that their sizes must be smaller than the original one, with emphasizing that the accepted solutions should cover the graph under test. The decision of accepting the updated solutions or rejecting them mainly is made in the greediness phase. Taking into consideration that the solutions are updating twice, as mentioned beforehand, the greediness phase is accomplished also twice, once is after the updating process in the individual competition phase as shown in lines 14-18. Once again is at the end of each iteration of the team competition phase as presented in lines 30-34.

The last phase that is included in the main phase is the elitism phase. In this phase, as recommended in [25], one-third of all solutions which have worst sizes (i.e. largest sizes) are replaced with these one-third solutions which have the best sizes (i.e. smallest sizes). With an aim to perform this replacement, first of all, the solutions from all groups must be collected back to the population *P*. In the aftermath, these solutions are sorted in a non-descending manner based on their sizes. As clearly observed in Algorithm 2, the elitism phase is implemented in lines 37-39. Ultimately, after executing all *M* iterations, the best-obtained solution is announced as *MVC* like is reveled in line 41.

It is worth noting that the duplicates removing phase is ignored during the adaptation of MVPA to the MVCP. Since it is inapplicable in case of the MVCP, this inapplicability attributed to that in some cases, the number of solutions of the graph under consideration is less than the initial population size. Thereupon, the duplication is unavoidable. Consequently, this phase cannot be applied.

V. Analytical Evaluation

This section offers a detailed discussion of an analytical evaluation for MVPA_MVCP algorithm in terms of the run time complexity. Given that, *M* is the maximum number of iterations, *I* is the initial population size, *T* is the number of groups, and |*V*| and |*E*| are, respectively, the number of vertices and edges in the graph under test.

The run time of an algorithm, as indicated in [15], is described as the number of steps executed over a particular size of input. The run time complexity calculated in this section is the average-case of the run time.

The run time complexity of MVPA_MVCP algorithm is the result of summing the run time complexity of its phases. Nonetheless, the run time complexity of creating the initial population solutions and calculating their sizes are not taken into consideration. This is due to the fact that it is assumed to be a preprocessing step. Theorem 1 remarks MVPA_MVCP algorithm average-case run time complexity. Yet, the details of calculating this complexity are illuminated in the following proof based on tracing the steps and phases listed in Algorithm 2.

Theorem 1 The average-case run time complexity of MVPA_MVCP algorithm is $\Theta\big(I \times (|V| + |E|)\big)$.

Proof: In the initial phase (lines 1-2), the process of generating the initial population is ignored. This is because it is assumed to be a preprocessing step. Finding the best solution in *P* requires *I* steps. Thus, the total run time complexity of the initial phase is *I*.

The main phase (lines 3-40) iterates *M* times. In each time, the following steps and phases are executed:

- Distributing the population solutions into *T* groups needs *I* steps.
- The competition phase (lines 5-36) iterates *T* times by executing the following steps and phases:

- Finding the best solution in any group requires $\frac{I}{T}$ steps. For the group that is currently processed and the randomly retrieved group, they require $2 \times \frac{I}{T}$ steps.
- The individual competition phase (lines 9-19), in this phase, processing all solutions of the group under consideration requires $\frac{I}{T}$ iterations. In each iteration, $|V|$ steps are needed to update the vertices of the processed solution. Another $|V|$ steps are needed to check the updated values of the vertices in the bound checking phase. To accomplish the greediness phase, and in order to decide on accepting the updated solutions or not, it is needed to check the updated solution capability of covering the graph under test. This checking entails dropping all edges of the solution vertices. The number of these edges may range from 1 to |*E*| as an upper limit. Therefore, the average number of steps that are maybe needed is $\frac{\sum_{i=1}^{|E|} i}{|E|} = \frac{|E| \times (|E|+1)}{2 \times |E|} = \frac{|E|+1}{2}$ steps. Consequently, the run time complexity of the individual competition phase, including the first executions of the bound checking phase and the greediness phase, is $\frac{I}{T} \times \left(2 \times |V| + \left(\frac{|E|+1}{2}\right)\right)$.
- The team competition phase (lines 20-35) follows the same main steps and phases included in the individual competition phase. It processes $\frac{I}{T}$ solutions. For each solution, all its vertices values are updated and checked with $2 \times |V|$ steps. Additionally, its capability of covering the graph under test is checked with $\frac{|E|+1}{2}$ steps, on average. As a result, the run time complexity of the team competition phase is also $\frac{I}{T} \times \left(2 \times |V| + \left(\frac{|E|+1}{2}\right)\right)$. Based on the above analysis, the total run time complexity of the competition phase is

$$T \times \left(2 \times \frac{I}{T} + 2 \times \left(\frac{I}{T} \times \left(2 \times |V| + \left(\frac{|E|+1}{2}\right)\right)\right)\right)$$

- The elitism phase (lines 37-39), the first step of applying this phase is to recollect the solutions from all the groups back to the population *P*. Actually, this step costs $T \times \frac{I}{T} = I$ steps. Worst one-third solutions that are replaced by the best one-third solutions costs $\frac{I}{3}$ steps. But to be able to do this replacement, sorting the solutions regarding their sizes is needed. Radix sort is chosen to perform this task. The time complexity of radix sort is $\Theta(I \times D)$, where D is the number of digits of the largest number to be sorted [14]. Accordingly, the elitism phase run time complexity is $I + (I \times D) + \frac{I}{3}$.

That is to say, the total run time complexity of the main phase is

$$M \times \left(I + \left(T \times \left(2 \times \frac{I}{T} + 2 \times \left(\frac{I}{T} \times \left(2 \times |V| + \left(\frac{|E| + 1}{2} \right) \right) \right) \right) \right) + \left(I + (I \times D) + \frac{I}{3} \right) \right)$$

In conclusion, the overall run time complexity of the MVPA_MVCP algorithm is

$$I + \left(M \times \left(I + \left(T \times \left(2 \times \frac{I}{T} + 2 \times \left(\frac{I}{T} \times \left(2 \times |V| + \left(\frac{|E| + 1}{2} \right) \right) \right) \right) \right) + \left(I + (I \times D) + \frac{I}{3} \right) \right) \right)$$

However, M, T and D can be dropped since they are constants which are much less than I, $|V|$, and $|E|$. As a result and by dropping all other constants, the average-case run time complexity of MVPA_MVCP algorithm is ended to be $\Theta(I \times (|V| + |E|))$. This completed the proof of Theorem 1.

VI. Experimental Results and Discussion

To evaluate the performance of the MVPA_MVCP algorithm experimentally, it is first implemented using Java programming language under 64-bit windows 10 pro operating system. Then, the MVPA_MVCP algorithm was tested on a Microsoft Azure virtual machine which has a 2.00 GHz Intel Xeon processor with 64 GB memory. Different instances of the DIMACS benchmark for the MVCP have been used to test the MVPA_MVCP algorithm. These instances are listed in Table II [31, 32], where the benchmark instances name, number of vertices, number of edges and their best known optimal solutions sizes [32] are presented.

As an overview of Microsoft Azure, it is one of the most recent services of Microsoft [33]. It is a cloud computing service that provides facilities for building, testing, deploying, and managing applications and services over Microsoft global datacenters network. In addition to supporting many frameworks and tools, it supports various programming languages. One of the most important facilities that Microsoft Azure implemented as a computer service is the creation of virtual machines [33].

TABLE. II. DIMACS Instances for Minimum Vertex Cover Problem

Benchmark Instances	Number of Vertices	Number of Edges	Best Known Optimal Solution Size
johnson8-2-4	28	168	24
graph50_6	50	857	38
graph50_10	50	612	35
Hamming6-4	64	1312	60
graph100_10	100	4207	70
johnson16-2-4	120	1680	112
keller4	171	5100	160
cfat200_1	200	18366	188
brock200_2	200	10024	188
Hamming8_4	256	11776	240
phat300_1	300	33917	292
phat300_2	300	22922	275
sanr400_0.5	400	39816	387
johnson32-2-4	496	14880	480
phat700-1	700	99800	689

In point of fact, the evaluation of the MVPA_MVCP algorithm is based on two performance metrics, the gained solution quality and the run time. For each benchmark instance, all tests are performed 10 times. The best-case, average-case and worst-case of these 10 tries are recorded for the solution quality metric. The run times of these three cases are also recorded. Besides, it is important to emphasize that the average-case of the solution quality metric is calculated as $\left\lceil \frac{\sum_{i=1}^{10} r_i}{10} \right\rceil$, where r_i is the minimum solution size gained in try i.

It is essential to draw the attention to the values of the main variables of the MVPA_MVCP algorithm before beginning to show and discuss the experimental outcomes. Regarding the initial population size, the number of groups, and k constant that is used in (2), they are specified as recommended in [25]. Their values are 100, 5, and 1, respectively. As regards the maximum number of iterations (M), several experiments have been performed to explore its best possible value that achieves a compromise between the gained solution quality and the run time. Three instances were used as samples for the conducted exploratory experiments. These instances varied in size; small (graph50_6, 50 vertices), medium (Hamming8_4, 256 vertices) and large (phat700-1, 700 vertices). The implemented MVPA_MVCP algorithm was executed using these three instances with M values 2, 4, 6, 8, and 10. The results showed that the most appropriate value was 6. This is because of the fact that the value of best-gained solutions of all three instances has not changed after the sixth iteration. On this foundation, M was assigned to 6 in all conducted experiments.

A. Solution Quality

First and foremost, solution quality is considered as one of the most expressive performance metrics to evaluate the metaheuristic algorithms. It specifies how much a gained solution has diverted from the optimal one. In the conducted experiments, the quality of a gained solution can be assessed since the best known optimal sizes of the solutions of the selected DIMACS data instances are recorded [32]. As an evaluator metric of the quality of gained solutions, the approximation ratio concept is used as in [21]. Mathematically, the approximation ratio is calculated as $\rho_\alpha = \frac{\alpha}{\beta}$, where ρ_α is the approximation ratio of the size of gained solution (α), and β is the size of the best known optimal solution. If the value of ρ_α equals to 1, then the size of the gained solution is the same as the size of the best known optimal solution. But with a value above 1, the size of the gained solution is worse than the size of the best known optimal solution.

In light of that, Table III demonstrates the best, average, and worst sizes of the obtained solutions and their respective approximation ratios. In this table, the chosen benchmark instances are sorted in non-descending order according to their number of vertices. Additionally, the noted bolded values appear in Table III indicate to those solutions that their sizes equal to the best known optimal solutions sizes. For the best case, average case and worst case, the MVPA_MVCP algorithm could gain respectively, seven, three and two solutions, with sizes equal to the best known optimal solutions sizes.

With respect to the approximation ratio, the gained solutions that have sizes equal to the sizes of the best known optimal solutions, their approximation ratios are equal to 1. Intuitively, the solutions that have sizes larger than the sizes of the best known optimal solutions, their values of the approximation ratios are greater than 1. All approximation ratios that have values equal to 1 are also bolded in Table III.

As an accumulated view, the last row of Table III shows the average of the approximation ratio values of all benchmark instances for all cases (i.e. best, average, and worst). These average values indicate that, on average, the MVPA_MVCP algorithm slightly diverted by only 0.01, 0.021, and 0.033 from the best known optimal solutions in the best case, average case and worst case, respectively.

B. Run Time

In order to discuss the run time (*RT*), the number of edges of a graph must be taken into consideration. Indeed, this number significantly impacts the entire *RT*, based on the fact that it impacts one process that is frequently repeated. Usually, this process checks the solution's capability to cover the involved graph. Actually, this check is performed by removing the edges of those vertices which are composing the solution. Thusly, when the number of edges becomes larger, more *RT* is needed to end the checking test. Consequently, the total *RT* will increase. Taking this fact into account, the DIMAC benchmark instances in Table IV are re-sorted in an ascending manner depending on their number of edges to clarify the effect of this number on the *RT*.

In Table IV, the *RT* (in seconds) of executing the MVPA_MVCP algorithm over the selected benchmark instances are recorded. The relationship between the number of edges and the *RT*, which outlined beforehand, can be clearly observed in all cases (best, average and worst) of gained solutions sizes. In fact, the general observation in Table IV is that, as the number of edges increases, the *RT* increases too.

Furthermore, Fig. 2 is created to graphically clarify the behaviour of the *RT* when the number of edges increases. Particularly, Fig. 2 demonstrates how long the average-case solutions can be accomplished. Moreover, since there is a large difference between the smallest and largest values of the *RT* in Table IV, the vertical axis of Fig. 2 is labelled by the logarithmic values of base 10 of the *RT*. This is to present these times more clearly. As laid out in Fig. 2, it also depicts the direct relationship between the *RT* and the number of edges.

TABLE. III. THE BEST, AVERAGE AND WORST GAINED SOLUTIONS SIZES (α) AND THEIR APPROXIMATION RATIOS (ρ_α)

Benchmark Instances	**Number of Vertices**	**Best Known Optimal Solution Size β**	***Best_α***	***Best_ρ_α***	***Average_α***	***Average_ρ_α***	***Worst_α***	***Worst_ρ_α***
johnson8-2-4	28	24	**24**	**1**	**24**	**1**	**24**	**1**
graph50_6	50	38	**38**	**1**	39	1.026	40	1.053
graph50_10	50	35	**35**	**1**	38	1.086	40	1.143
Hamming6-4	64	60	**60**	**1**	61	1.017	62	1.033
graph100_10	100	70	72	1.029	73	1.043	76	1.086
johnson16-2-4	120	112	**112**	**1**	**112**	**1**	113	1.009
keller4	171	160	162	1.013	164	1.025	165	1.031
cfat200_1	200	188	**188**	**1**	**188**	**1**	**188**	**1**
brock200_2	200	188	191	1.016	192	1.021	193	1.027
Hamming8_4	256	240	248	1.033	248	1.033	249	1.038
phat300_1	300	292	**292**	**1**	293	1.003	294	1.007
phat300_2	300	275	285	1.036	285	1.036	287	1.044
sanr400_0.5	400	387	393	1.016	393	1.016	395	1.021
johnson32-2-4	496	480	482	1.004	482	1.004	483	1.006
phat700-1	700	689	694	1.007	694	1.007	696	1.01
		Average of ρ_{GSS}		1.01		1.021		1.033

TABLE. IV. THE RUNTIME (*RT*) (IN SECONDS) OF THE TESTED BENCHMARK INSTANCES

Benchmark Instances	**Number of Edges**	RT_{Best_α}	$RT_{Average\_\alpha}$	RT_{Worst_α}
johnson8-2-4	168	0.021	0.021	0.025
graph50_10	612	0.109	0.109	0.094
graph50_6	857	0.2	0.28	0.213
Hamming6-4	1213	0.295	0.262	0.239
johnson16-2-4	1680	0.587	0.532	0.445
graph100_10	4207	1.469	1.689	1.407
keller4	5100	1.9	2.218	2.201
brock200_2	10024	4.995	5.271	6.457
Hamming8_4	11776	8.772	9.176	8.448
johnson32-2-4	14880	10.75	19.622	12.687
cfat200_1	18366	15.806	16.106	16.13
phat300_2	22922	20.02	24.803	22.758
phat300_1	33917	28.682	27.965	27.077
sanr400_0.5	39816	38.961	60.254	38.665
phat700-1	183651	133.538	115.903	61.814

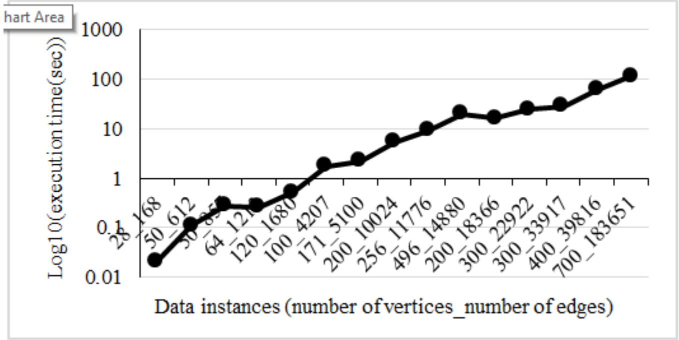


Fig. 2. The Run Time of Gaining Average-Case Solutions.

VII. CONCLUSIONS AND FUTURE WORK

On one side, the MVCP is one of the NP-hard problems that many scientists have been dealing with. This is because it has demonstrated its flexibility in solving problems in several applications in real-life. On the other side, the MVPA has recently developed as one of the metaheuristic algorithms that has been influenced by its concept in team sports. In this paper, the MVPA_MVCP algorithm is presented as an adaptation of the MVPA for the MVCP. The MVPA_MVCP algorithm is analytically evaluated, and several tests are conducted with a target of experimental evaluation. Regarding the analytical evaluation, the MVPA_MVCP algorithm is evaluated in terms of the run time complexity. It has been shown that its average-case run time complexity ended to be $\Theta(I(|V| + |E|))$, where I is the size of the initial population, $|V|$ is the number of vertices and $|E|$ is the number of edges of the graph under test.

For the conducted experiments, the MVPA_MVCP algorithm is developed using Java programming language and it is executed on a Microsoft Azure virtual machine that has a 2.0 GHz Intel Xeon processor with 64 GB memory. As test data set, 15 DIMACS benchmark instances for minimum vertex cover problem are used.

The experimental results are evaluated in terms of the run time; in addition to the quality of the gained solutions. These results clarified that there is a direct relation between the number of edges of the processed graph and the run time. Where when the number of edges increases, the run time increases too. Besides, they showed that, in the best case, the MVPA_MVCP algorithm could gain seven solutions that have sizes exactly as the best known optimal solutions sizes.

As future work, the MVPA_MVCP algorithm can be compared with other metaheuristic algorithms such as GA and ACO. Microsoft Azure service of creating multi-core virtual machines can be also invested to parallelize the MVPA_MVCP algorithm. Additionally, it can be parallelized over some types of interconnection networks like Chained-Cubic Tree interconnection network (CCT) [34], Optical Chained-Cubic Tree interconnection network (OCCT) [35], and Optical Transpose Interconnection System (OTIS) networks; such as OTIS-hypercube, OTIS-mesh, OTIS hyper hexa-cell, and OTIS mesh of trees [36, 37]. These interconnection networks exposed their usefulness for solving various problems in a parallel mode [36-38]. Solving the MVPA_MVCP algorithm on parallel computing environment could greatly reduce the run time, and it should not affect the quality of the obtained solutions.

REFERENCES

[1] Y. Raju, and N. Devarakonda, "Cluster based Hybrid Approach to Task Scheduling in Cloud Environment," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 10, no. 4, pp. 425–429, 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100452.

[2] A. Alazzawi, H. Rais, and Sh. Basri, "ABCVS: An Artificial Bee Colony for Generating Variable T-Way Test Sets," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 10, no. 4, pp. 259-274, 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100431.

[3] I. Sabbani, B. Omar, and D. Eszetergar-Kiss, "Simulation Results for a Daily Activity Chain Optimization Method based on Ant Colony Algorithm with Time Windows," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 10, no. 1, pp. 425-430, 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100156.

[4] A. Wicaksono, and A. Supianto, "Hyper Parameter Optimization using Genetic Algorithm on Machine Learning Methods for Online News Popularity Prediction," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 9, no. 12, pp. 263-267, 2018. http://dx.doi.org/10.14569/IJACSA.2018.091238.

[5] D. Poole, and A. Mackworth, Artificial Intelligence: Foundations of Computational Agents, 2nd ed., Cambridge University Press, 2017.

[6] R. Masadeh, B. Mahafzah, and A. Sharieh, "Sea lion optimization algorithm," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 388–395, 2019.

[7] R. Masadeh, A. Sharieh, and B. Mahafzah, "Humpback whale optimization algorithm based on vocal behavior for task scheduling in cloud computing," International Journal of Advanced Science and Technology, vol. 13, no. 3, pp. 121–140, 2019.

[8] M. Alshraideh, E. Jawabreh, B. Mahafzah, and H. AL Harahsheh, "Applying genetic algorithms to test JUH DBs exceptions," International Journal of Advanced Computer Science and Applications, vol. 4, no. 7, pp. 8–20, 2013.

[9] M. Alshraideh, B. Mahafzah, H. Eyal Salman, and I. Salah, "Using genetic algorithm as test data generator for stored PL/SQL program units," Journal of Software Engineering and Applications, vol. 6, no. 2, pp. 65–73, 2013.

[10] U. Chakraborty, D. Konar, and C. Chakraborty, “A GA based Approach to Find Minimal Vertex Cover,” International Journal of Computer Applications (IJCA), National Conference cum Workshop on Bioinformatics and Computational Biology, NCWBCB, vol. 3, pp. 5–7, 2014.

[11] H. Bhasin and M. Amini, “The applicability of genetic algorithm to vertex cover,” International Journal of Computer Applications, vol. 123, no. 17, pp. 29-34, 2015.

[12] A. Pat, "Ant colony optimization and hypergraph covering problems," IEEE Congress on Evolutionary Computation (CEC), Beijing, China, July 6-11, 2014.

[13] R. Jovanovic and M. Tuba, “An ant colony optimization algorithm with improved pheromone correction strategy for the minimum weight vertex cover problem,” Applied Soft Computing, vol. 11, no. 8, pp. 5360–5366, 2011. https://doi.org/10.1016/j.asoc.2011.05.023.

[14] Z. Guangyong, X. Yuimg, L. Kenli, and S. Shibing, “Hybrid chemical reaction optimization algorithm for minimum vertex cover problem,” Electronic Technology and Information Science, vol. 33, no. 9, 2016.

[15] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms, 3rd ed., The MIT Press, 2009.

[16] V. Kavalci, A. Ural, and O. Dagdeviren, “Distributed vertex cover algorithms for wireless sensor networks,” International Journal of Computer Networks and Communications, vol. 6, no. 1, pp. 95–110, 2014.

[17] M. Barah and E. Mazaheri, Matching Theory. In: Farahani R, Miandoabchi E (ed) Graph Theory for Operations Research and Management: Applications in Industrial Engineering. Portland, Book News Inc., pp 127–141, 2012.

[18] T. Can, Introduction to Bioinformatics. In: Yousef M., Allmer J. (eds) miRNomics: MicroRNA Biology and Computational Analysis. Methods in Molecular Biology (Methods and Protocols), vol 1107, pp. 51-71. Humana Press, Totowa, NJ, 2014. https://doi.org/10.1007/978-1-62703-748-8_4.

[19] H. Moser, Exact algorithms for generalizations of vertex cover. Master thesis, Friedrich-Schiller University, Jena, Germany, 2005.

[20] G. Kochenberger, M. Lewis, F. Glover, and H. Wang, “Exact solutions to generalized vertex covering problems: A comparison of two models,” Optim. Lett., vol. 9, no. 7, pp. 1331–1339, 2015. https://doi.org/10.1007/s11590-015-0851-1.

[21] I. Khan and S. Khan, “Experimental comparison of five approximation algorithms for minimum vertex cover,” International Journal of u-and e-Service, vol. 7, no. 6, pp. 69–84, 2014.

[22] M. Eshtey, A. Sliet, and A. Sharieh, “NMVSA greedy solution for vertex cover problem,” International Journal of Advanced Computer Science and Applications, vol. 7, no. 3, pp. 60–64, 2016.

[23] S. Cai, K. Su, and A. Sattar, “Local Search with edge weighting and configuration checking heuristics for minimum vertex cover,” Artificial Intelligence, vol. 175, pp. 1672–1696, 2011. https://doi.org/10.1016/j.artint.2011.03.003.

[24] S. Balachandar and K. Kannan, “A Meta-heuristic algorithm for vertex covering problem based on gravity,” International Journal of Mathematical and Statistical Sciences, vol. 1, no. 3, pp. 130–136, 2009.

[25] H. Bouchekara, “Most valuable player algorithm: a novel optimization algorithm inspired from sport,” Oper. Res. Int. J., 2017. https://doi.org/10.1007/s12351-017-0320-y.

[26] Sh. Cai, K. Su, C. Luo, and A. Sattar, “NuMVC: an efficient local search algorithm for minimum vertex cover,” Journal of Artificial Intelligence Research, vol. 46, pp. 687–716, 2013.

[27] Z. Fang, Y. Chu, K. Qiao, X. Feng, and K. Xu, “Combining edge weight and vertex weight for minimum vertex cover problem,” Proceedings of International Workshop on Frontiers in Algorithmics, vol. 1, Zhangjiajie, China, pp. 71–81, June 2014.

[28] H. Bouchekara, A. Orlandi, M. Al-Qda, and F. Paulis, “Most valuable player algorithm for circular antenna arrays optimization to maximum sidelobe levels reduction,” IEEE Transactions on electromagnetic compatibility, vol. 60, no. 6, pp. 1655-1661, 2018.

[29] B. Alatas, “Sports inspired computational intelligence algorithms for global optimization,” Artif. Intell. Rev., 2017. https://doi.org/10.1007/s10462-017-9587-x.

[30] S. Luke, Essentials of Metaheuristics. Lulu Publisher, 2013. http://cs.gmu.edu/~sean/book/metaheuristics/

[31] D. Johnson and M. Trick, Cliques, Coloring and Satisfiability: Second DIMACS Implementation Challenge, DIMACS Series, American Mathematical Society, Providence RI, 26, 1996.

[32] F. Mascia, DIMACS benchmark set, 2015. http://iridia.ulb.ac.be /~fmascia/maximum_clique/DIMACS-benchmark. Accessed 1 June 2019.

[33] M. Collier and R. Shahan, Fundamentals of Azure. Microsoft Press, Redmond, Washington, 2016.

[34] M. Abdullah, E. Abuelrub, and B. Mahafzah, “The chained-cubic tree interconnection network,” International Arab Journal of Information Technology, vol. 8, no. 3, pp. 334–343, 2011.

[35] B. Mahafzah, M. Alshraideh, T. Abu-Kabeer, E. Ahmad, and N. Hamad, “The optical chained-cubic tree interconnection network: topological structure and properties,” Computers & Electrical Engineering, vol. 38, no. 2, pp. 330–345, 2012. https://doi.org/10.1016/ j.compeleceng.2011.11.023

[36] A. Al-Adwan, B. Mahafzah, and A. Sharieh, “Solving traveling salesman problem using parallel repetitive nearest neighbor algorithm on OTIS-Hypercube and OTIS-Mesh optoelectronic architectures,” Journal of Supercomputing, vol. 74, no. 1, pp. 1–36, 2018. https://doi.org/10.1007/s11227-017-2102-y

[37] A. Al-Adwan, A. Sharieh, and B. Mahafzah, “Parallel heuristic local search algorithm on OTIS hyper hexa-cell and OTIS mesh of trees optoelectronic architectures,” Applied Intelligence, vol. 49, no. 2, pp. 661–688, 2019. https://doi.org/10.1007/s10489-018-1283-2

[38] S. Baddar and B. Mahafzah, “Bitonic sort on a chained-cubic tree interconnection network,” Journal of Parallel and Distributed Computing, vol. 74, no 1, pp. 1744–1761, 2014. https://doi.org/10.1016/j.jpdc.2013.09.008.

Non-linear Dimensionality Reduction-based Intrusion Detection using Deep Autoencoder

S. Sreenivasa Chakravarthi[1]
Asst. Prof. of CSE, Center for Intelligent Computing
Sree Vidyanikethan Engineering College &
Research Scholar, SCSE, VIT, Chennai, India

R. Jagadeesh Kannan[2]
Professor and Dean
SCSE, VIT
Chennai, India

***Abstract*—The intrusion detection has become core part of any network of computers due to increasing amount of digital content available. In parallel, the data breaches and malware attacks have also grown in large numbers which makes the role of intrusion detection more essential. Even though many existing techniques are successfully used for detecting intruders but new variants of malware and attacks are being released every day. To counterfeit these new types of attacks, intrusion detection must be designed with state of art techniques such as Deep learning. At present the Deep learning techniques have a strong role in Natural Language Processing, Computer Vision and Speech Processing. This paper is focused on reviewing the role of deep learning techniques for intrusion detection and proposing an efficient deep Auto Encoder (AE) based intrusion detection technique. The intrusion detection is implemented in two stages with a binary classifier and multiclass classification algorithm (dense neural network). The performance of the proposed approach is presented and compared with parallel methods used for intrusion detection. The reconstruction error of the AE model is compared with the PCA and the performance of both anomaly detection and the multiclass classification is analyzed using metrics such as accuracy and false alarm rate. The compressed representation of the AE model helps to lessen the false alarm rate of both anomaly detection and attack classification using SVM and dense NN model respectively.**

Keywords*—Autoencoder; deep learning; principal component analysis; dense neural network; false alarm rate

I. Introduction

For detecting various security attacks and breaches with a network Intrusion Detection Systems (IDS) are essential tools. An ID system monitors the traffic in the network (both incoming and outgoing traffic bounds) and performs analysis to raise an alarm if there is anomaly being detected. Based on the approach used for intrusion detection, it can be classified either as signature based or anomaly based method [1]. The signature based intrusion detection method work by matching predefined rules against the pattern in the current network traffic and classifies it as intrusion if the pattern deviates from the normal pattern. It is effective for identifying known type of attacks and yields high accuracy in detecting and low false alarm rate. This approach cannot be used for detecting the new category of attacks as the predefined rules cannot match the unknown patterns in them. The anomaly based intrusion detection systems are capable of classifying even unknown or new category of attacks [2]. As per the theoretical proofs given in many literature the anomaly based detection are more accurate but in practice they suffer from high false alarm rate. Among the several challenges in detecting the intrusions two primary challenges include selection of optimal feature from the network traffic dataset for classification, and unavailability of supervised (labeled) dataset for training the machine learning models [3]. As the attacks patterns are changing over a period of time the set of features used for discriminating the attacks cannot be suitable for effective discrimination of new category of attacks [4].

It is inferred from various literature that machine learning techniques such as Artificial neural Networks, Support Vector Machine (SVM), Naive Byes Classifier, Random Forest (RF), and Self-Organizing Maps (SOM) have been utilized for developing an anomaly based intrusion detection approach. In general the anomaly based classifiers are trained to discriminate between the normal and anomalous traffic. In apart from the training process most of the anomaly based methods adopts a feature selection task to select an optimal set of features to better discriminate the anomalous traffic. During the feature selection process the high dimensional training dataset is reduced in to low dimensional representation and the redundant features are eliminated. For selecting an optimal set of features various methods including Genetic Algorithm, meta-heuristic algorithms, and Principal Component Analysis (PCA) are used [5]. This paper utilizes a deep auto-encoder for finding a compact representation of the input data and a dense neural network for classification of anomalous traffic.

The estimation of compact representation of the network traffic data is a preprocessing task before classifying it. Training a classifier on low dimensional input is much faster than the original input data. This process can also be considered as a dimensionality reduction step which has a regularization effect and prevents over-fitting. The other dimensionality reduction strategies are not effective when compared to AE based approach. When PCA or Zero Component Analysis (ZCA) is used for dimensionality reduction the memory requirements will be high if the feature space is having high dimension. The AE maps the original input data to a compact representation after sending data through two un-supervised training stages. The AE is a generative model and is able to learn and discover the semantic similarity and correlation among the input features [6, 7].

The remaining sections of this paper are structures as follows. Section II of this paper briefs some of the important works relevant to machine learning or deep learning based network intrusion detection approaches. In Section III, the architecture of the AE model and the procedure for its pre-training and training are presented. The efficiency of the proposed methodology for dimensionality reduction and classification (binary and multiclass classification) are presented in Section IV. Sub-section IV-A elaborates the modeling setup, and Subsection IV-B presents the analysis of the results. Finally, Section V concludes the paper.

II. Related Works

At present the role of Deep Learning can be found in many real time applications including intrusion detection and malware analysis. This section will present a detailed review of deep learning and its application in network intrusion detection. In [8] the authors have conducted a literary and experimental comparison between conventional network intrusion detection methods and deep learning based intrusion detection methods. Their experimental results proved that deep learning based detection techniques offered an improved performance when compared to their traditional counterparts. The problem of imbalanced dataset available for training the detection models can be overcome by using the oversampling technique SMOTE - Synthetic Minority Oversampling Technique [9]. For detecting the malware a stacked denoising AE was used in [10]. Also the same AE based classifier was used to classify executable files [11]. This model used API calls as the features for discriminating the malware from the normal codes. In another approach AE and Restricted Boltzmann Machine was combined used to detect malware using static and dynamic features [12]. In [13] Recurrent Neural Network based AE was developed to extract the features to best describe the malware from raw API calls. The RNN was trained with compact representation from the AE. The deep learning research community has also addressed the problem of outlier detection. The models developed vary on the structure, the application context, and motivation behind the opted strategy [14]. The work proposed in [16] introduced the usage of variation AE for detecting intruders. It was observed that the variation AE performed well in intrusion detection and also in outlier detection. A combination of hybrid AE along with Density Estimation (DE) model was used for detecting anomalies. The model is developed based on the density estimation of the compressed hidden layer representation of AE.

The model was constructed based on estimation of the density among the zipped hidden-layer notation of the applied AE. In another similar research work [17] the authors have used hybrid de-noising AEs in a stacked architecture. Their model utilized hex-based representation of portable executable files, without disassembling. The model training process does not include feature selection or extraction before processing the data which may induce the efficiency of the model. The proposed methodology provided a better discrimination between a legitimate network flow and an anomalous flow. The model output a fixed length vector for both detecting malware and classifying the attacks.

In [15] a three layered RNN architecture was used which utilizes 41 features as inputs and was able to detect intrusion out of four categories. The authors have not studied the performance of the model for binary classification. The nodes of the hidden layers are connected partially and hence it does not have the ability to model high dimensional features.

III. Background Concepts

This section presents some of the background concepts related to the techniques and algorithms used in this study. Deep Autoencoders follows unsupervised learning methodology for learning representation by using dense neural network architecture. The neural network is designed in such a way that the network yields a compressed representation of the input given to it. In a highly non-correlated input data the reconstruction of the input data from the compressed representation is a complex task. If there is a pattern exist in between the data then the pattern can be learned by passing the data through the layers of the neural architecture. Thus the network accepts an unlabeled set of samples and gives xˆ as output which is a reconstructed from the compressed representation of the given input. The network is trained to reduce the error in reconstruction denoted as $L(x, \hat{x})$. The last layer of the encoder decides the size of the data which pass through the decoder layer. The Principal Component Analysis (PCA) achieves a linear dimensionality reduction which is similar to the compressed representation of AE with linear activation function at each layer [18]. For detecting the intrusion patterns in the network data the model must be sensitive to the input for building an accurate reconstruction and at the same time it should be insensitive to the inputs that over-fit the model.

For achieving the above mentioned constraints the network is designed with a loss function with two terms. The first terms makes model to be sensitive to the inputs and the second term avoid the model from overfitting the training data. The alpha scaling parameter controls the trade-off between the two stated objectives. The AE can be considered as a nonlinear generalization of PCA and it is capable of learning non-linear relationship between the input data. Fig. 1 presents the schematic view of the PCA and AE based dimensionality reduction.

$$L(x, \hat{x}) + (\alpha * regularizer) \tag{1}$$

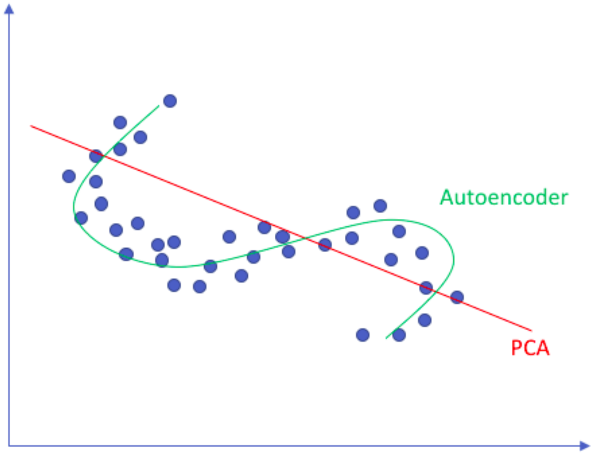


Fig. 1. Linear Versus Non-Linear Dimensionality Reduction.

When a higher dimensional data has to be classified then the AE can be used to estimate a lower dimensional representation of the data and the same can be decoded in to the original input data. The AE is used as a non-linear feature extractor which can be considered as non-linear generalization of Principal Component Analysis (PCA). It performs both encoding of the given input and decoding of the encoded i.e. the compressed feature representation. The encoded data are the dimensionally reduced features. In this study the encoded features are used for identifying anomaly in the network and as well the category of the attack.

The sparsity constraint can be imposed in the network by adding L1 regularization term to the loss function for penalizing the value of activation in the hidden layer along with a tuning parameter λ.

$$L(x, \hat{x}) + \lambda \sum_i \left| a_i{}^{(h)} \right| \quad (2)$$

IV. Proposed Methodology

The AE is constructed using encoder and decoder which do the transformation of input from the high dimensional space in to a compresses representation in a low dimensional space and reconstruction of the supplied input to the encoder respectively. The architecture of the AE model is presented in Fig. 2. During the training process the weights of the layers in the decoder and encoder are tuned sequentially. The objective is to minimize the reconstruction error. For training the AE model the back propagation learning algorithm is used and the layer weights are assigned randomly and from various experimental results presented in literature it is obvious that the selection of hyper-parameters play an important role in meeting the above stated objective.

For a given set of training examples *X = {x₁, x₂, x₃, x₄,……,xₘ}* where x_i is a d - dimensional feature vector. The encoding layers of the AE maps the d - dimensional input vector to a hidden vector representation h_i and the mapping function is denoted as f_θ and it is defined as in Eq. 3

$$h_i = f_\theta(x_i) = s(Wx_i + b) \quad (3)$$

Where 's' is a sigmoid activation denoted as

$$s(t) = \frac{1}{1 + exp^{-t}}$$

The decoding layers reconstruct the input data represented as vector y_i in *d*-dimensional space.

$$y_i = g_{\acute{\theta}}(x_i) = s(\acute{W} h_i + \acute{b}) \quad (4)$$

The aim of the training is to minimize the error between input and reconstructed output. The output of the training process is the optimal parameter $\theta$and $\acute{\theta}$. The error is estimated using a loss function and it is denoted as

$$L(x, y) = \frac{1}{m} \sum_{i=1}^{m} \|x_i - y_i\|^2 \quad (5)$$

The number of nodes in the code layer is a *hyper-parameter* that is initialized set before training the AE. The other hyper-parameters include number of layers, number of nodes per layer, and loss function, batch size and dropout rate. Either *mean squared error (MSE)* or *binary cross-entropy can be used depending on the input*. If the input values are in the range [0, 1] then typically cross-entropy can be used, otherwise mean squared error. As the value of hyper-parameters has a direct influence on the success of the training process, an optimal set of hyper-parameters is estimated using evolutionary optimization approach i.e. using Genetic Algorithm. During the hyper-parameter optimization the space of hyper-parameter is searched and an optimal set of parameters are found. Initially a random tuple of hyper-parameters are generated and the fitness value is estimated. Then the hyper-parameters tuples are ranked by their respective fitness values. The hyper-parameter tuples with low fitness values are replace with new hyper-parameter tuples generated by crossover and mutation functions. This process is repeated iteratively until there is no change in the fitness value for a set of consecutive iterations.

While optimizing the hyper-parameters using genetic algorithm; K-fold cross validation error is considered as the fitness function. As per K-fold cross validation for assessing the performance of the AE over a set of selected hyper-parameters a part available dataset is used for training the model and the remaining for testing it. The cross validation error can be estimated as follows.

$$CV(\hat{f}) = \frac{i}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-k(i)}(x_i)) \quad (6)$$

where $\hat{f}^{-k(i)}$denotes the model fitted with kth part of the data removed.

After training an AE model the network data from the data set is applied to it and compressed feature vector is used for training both a binary classifier and multi-classifier for anomaly and specific intrusion type detection as presented in Fig. 3.

For binary classification the Support Vector Machine (SVM) model with a Radial Basis Function (RBF) kernel function was utilized and a Dense Neural Network was used for Multi-Classification.

$$K(X, X') = \exp[-\gamma \|x - x'\|^2] \quad (7)$$

The RBF kernel helps estimating the similarity between vectors by measuring the squared Euclidean distance between them. When the two vectors are close together then the value $\|x - x'\|$ will be small. As the value of γ > 0, it is obvious that $-\gamma\|x - x'\|^2$ will also be larger. Hence vectors appearing closer will have a large RBF kernel value when compared to vectors which are separated by a nominal distance. For evaluating the performance of the multi-classifier, various models including Naïve Bayes & Random Forest are utilized for classifying the category of attack. The SVM and the dense NN are trained with the compressed representation of the input.

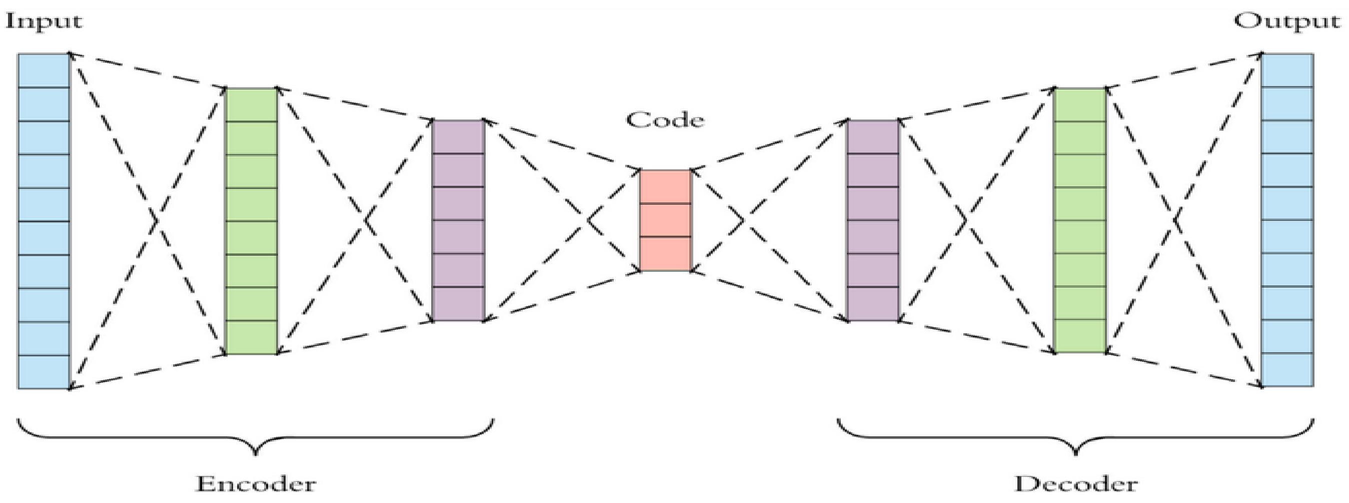


Fig. 2. Schematic View of Autoencoder Architecture.

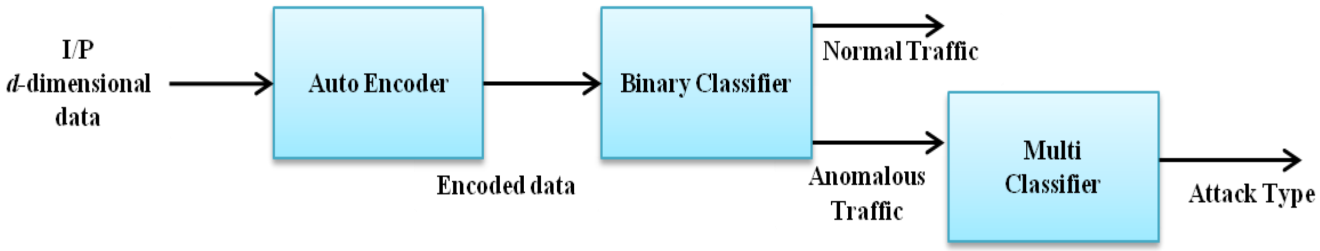


Fig. 3. Block Diagram of Proposed Intrusion Detection.

V. Experiments and Results

A. Dataset

The various attack type detected using signature based approach using a deep neural network is discussed in this section. The dataset used for training and evaluation of the intrusion detection model is obtained from UNSW dataset [19]. The dataset is constructed using the packets generated from the IXIA Perfect Storm tool for both normal and various attack categories possible in the cloud network. The TCP logs were used to extract the relevant attributes which better represents the attacks. Using k-fold cross validation the dataset is divided for training and testing the detection model. Description about the various attacks along with normal data is summarized in Table I.

As discussed in previous section the AE is trained in unsupervised mode and the trained model is used for constructing a compact representation of the input. The Compressed data is used for training the SVM and other multi classifier models. While training the AE the hyper-parameters are estimated using Genetic Algorithm. The error in the reconstruction of the given input to the AE model is analyzed to measure the performance of the AE model.

The numeric features are normalized for removing the effect of original feature value scales. Each feature is normalized and the normalized data is represented as

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \qquad (8)$$

TABLE I. Dataset Records Distribution

Type	**No. of Records**	**Description**
Normal	2,218,761	Normal legitimate transaction data.
Fuzzers	24,246	Cause a service suspension by sending in a sequence of randomly generated data.
Analysis	2,677	This includes various attacks of scanning the communication ports, spam and penetration of html files in to cloud nodes.
Backdoors	2,329	Unauthorized access to virtual machine and the data stored is gained bypassing the security mechanism by stealth attack.
DoS	16,353	The service is made unavailable to authorized users by disrupting the host connected via Internet.
Exploits	44,525	Exploits the knowledge and information on a known security issue in an operating system of a virtual machine.
Generic	215,481	Irrespective of the structure of the block cipher a common technique against all block ciphers are used with a given block and key length.
Reconnaissance	13,987	Includes all strikes that are capable of collecting information by simulating the attacks.
Shellcode	1,511	The vulnerability in the software is exploited by a small piece of code as the pay-load.
Worms	174	The attack spread itself by replicating to other virtual machines in the cloud environment. Often, it uses a computer network to spread itself, relying on security failures on the target computer to access it.

After normalization all numeric features will be ranged between 0 and 1. The objective of the AE is to minimize the reconstruction error. The training process tunes the weights connecting the layer of the AE model. If the values of the weights are tuned to higher values then it make the generated features more dependent on the structure of the network; but not on the input. In order to avoid this dependency of features on the network structure weight decay regularization is imposed to maintain the weights of the network smaller. Thus the objective function can be defined as

$$L(x,y) = \frac{1}{m}\sum_{i=1}^{m}\|x_i - y_i\|^2 + \lambda\,\|W\|^2 \qquad (9)$$

where $\|W\|^2$ is the weight decay regularization term to ensure a smaller value of weights tuned during the training process. The parameter λ scales the regularization term and can be considered as a hyperparameter of the AE model.

B. Results

Fig. 4 presents the performance of the AE model with a plot of average reconstruction error of PCA and AE for different encoding dimension.

The AE model performs better than PCA at reconstructing the input data set when the number of dimension in encoded output is small, but the error converges as the dimension increases. For very large data sets this difference will be larger and means a smaller data set could be used for the same error as PCA. When dealing with big data this is an important property.

Based on the previous results and comparison with the PCA encoding dimension size is selected as 15 and the value of λ – the weight regularization term is fixed as 3×10^{-3}.There are no guidelines to choose the size of the bottleneck layer in the AE unlike PCA. With PCA, the top $\boldsymbol{k}$ components can be chosen to factor. PCA is used as a guide to choose the value of k - the encoding dimension size.

The performance of the anomaly detection is analyzed based on the following metrics Detection Rate (DR), Accuracy (ACC), and False Alarm Rate (FA).

$$DR = \frac{TP}{(TP+FN)} \qquad (10)$$

$$ACC = \frac{TP+TN}{No.of\ samples} \qquad (11)$$

$$FA = \frac{FP}{FP+TN} \qquad (12)$$

where TP, TN, FP, and FN are True Positive, True Negative, False Positive, and False Negative values of the anomaly detection using the SVM. The training loss trends as depicted in Fig. 5 and the plot in Fig. 6 exhibits the sync between the training and validation loss which proves that the model is not overfitting.

To avoid overfitting of the model is trained with a dropout rate of 0.3 for regularization. The dropout rate is chosen using the genetic algorithm based hyper-parameter optimization technique along with other hyper-parameters.

After completing the training and validation of the AE model then the encoded output is used for training SVM model used as binary classifier for detecting the anomalies in the network. As the accuracy measure alone is not sufficient for analyzing the performance, the Receiver Operating Characteristics curve is plotted to analyze the trade-off between the false positive rate and the true positive rate of the binary classifier and the RoC curve, as presented in Fig. 7, is preferred when the training samples are balanced between the normal and anomalous samples. There are a number of methods available to oversample a dataset used in a typical classification problem. The most common technique for oversampling the imbalanced class in a training set and the technique in the experiments is SMOTE: Synthetic Minority Over-sampling Technique. For a set of training set with ***m*** samples, and ***n*** features in the feature space of the data. To then oversample, take a sample from the dataset, and consider its *k* nearest neighbors (in feature space). To create a synthetic data point, take the vector between one of those *k* neighbors, and the current data point. Multiply this vector by a random number *x* which lies between 0, and 1. Add this to the current data point to create the new, synthetic data point. The area under the curve is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). The area under curve is given by.

$$A = \int_{x=0}^{1} TPR\left(FPR^{-1}(x)\right)dx \qquad (13)$$

The area under the roc curve is estimated as 0.82 which proves that the accuracy of the anomalous detection is high. The Random Forest classifier and Naïve Bayes classifiers were also trained with the dimensionality data generated by the AE model and their results were compared with the results of the deep NN. Fig. 8 presents the plot results obtained from other classifiers and deep NN.

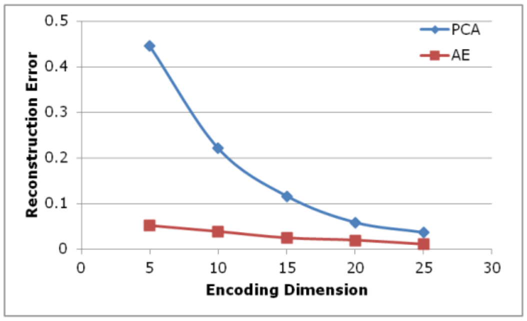


Fig. 4. Comparison of Reconstruction Error between PCA and AE.

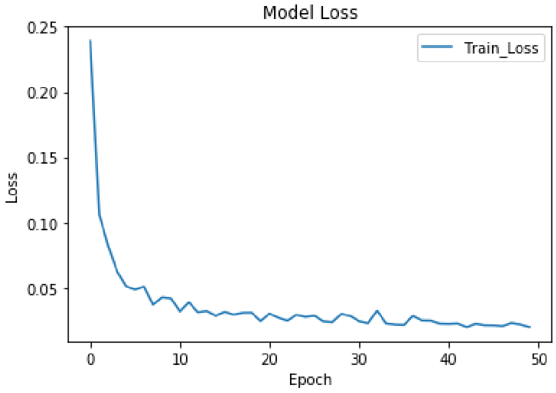


Fig. 5. Autoencoder training loss curve (λ = 0.003).

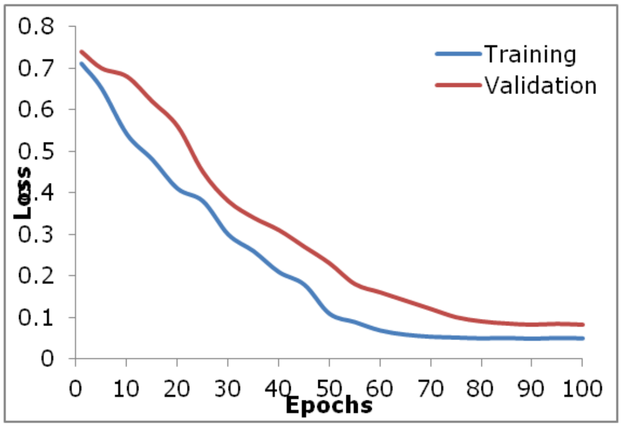


Fig. 6. Autoencoder Training History.

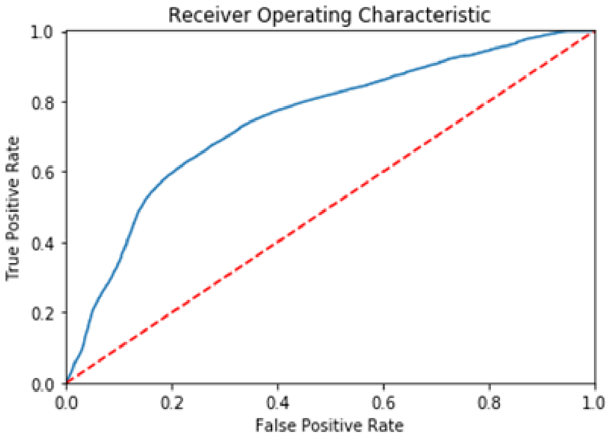


Fig. 7. RoC Curve for Anomaly Detection.

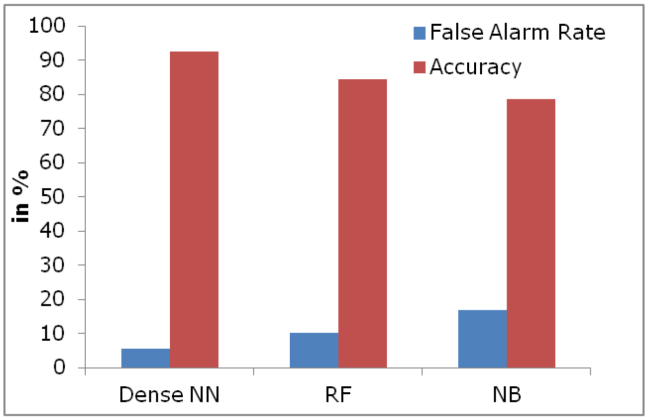


Fig. 8. Performance Comparison of Attack Classification.

VI. Conclusion

This paper presented an intrusion detection using autoencoder trained to generate a compact representation of the input data. The strength of the AE model complements the anomaly detection and attack classification process. The AE model is trained with a set of optimal hyper-parameters estimated using an optimization technique. During training the AE model is prevented from overfitting by using a regularization term and dropout at the hidden layers.

The performance of the AE model is measured using the reconstruction error and compared with the PCA. The accuracy of the intrusion detection process is implemented in a two level. The reduced feature set is applied to the SVM model for detecting the anomaly in the network and subsequently the same data sample is supplied through a dense network if there is an anomaly detected in the network. This methodology is quite simple to implement in a real time environment provided ample of data available for training the binary and multiclass classifiers.

The feasibility and effectiveness of the proposed anomaly detection model was evaluated using ROC curve. Future work will focus on designing a middleware to detect the anomalous traffic in the network in real time using deep learning techniques. This work validated the ability of the deep AE model in reconstructing the data points drawn from different distributions. The overall performance of the intrusion detection on the raw input data, the PCA transformed data, and the autoencoder generated data are compared and the AE seems to be the reason for improvement in accuracy of the attack detection (multi-class classification) especially.

References

[1] Casas, Pedro, Johan Mazel, and Philippe Owezarski. "Unsupervised network intrusion detection systems: Detecting the unknown without knowledge." Computer Communications vol.35, no.7, pp. 772-783, April 2012.

[2] Liao, Hung-Jen, et al. "Intrusion detection system: A comprehensive review." Journal of Network and Computer Applications, vol.36 no.1, pp. 16-24, January 2013.

[3] Bhuyan, Monowar H., Dhruba Kumar Bhattacharyya, and Jugal K. Kalita. "Network anomaly detection: methods, systems and tools." IEEE Communications Surveys & Tutorials vol.16, no.1, pp. 303-336, 2014.

[4] Heba, F. Eid, et al. "Principle components analysis and support vector machine based intrusion detection system." 10th International Conference on Intelligent Systems Design and Applications. IEEE, December 2010.

[5] George, Annie, and A. V. Vidyapeetham. "Anomaly detection based on machine learning: dimensionality reduction using PCA and classification using SVM." International Journal of Computer Applications, vol.47, no.21, pp. 5-8, June 2012.

[6] Shone, Nathan, et al. "A deep learning approach to network intrusion detection." IEEE Transactions on Emerging Topics in Computational Intelligence vol.2, no.1, pp. 41-50, February 2012.

[7] Mahmood Yousefi-Azar ; Vijay Varadharajan ; Len Hamey ; Uday Tupakula, "Autoencoder-based feature learning for cyber security applications." International Joint Conference on Neural Networks (IJCNN). IEEE, May 2017.

[8] Yin, Chuanlong, et al. "A deep learning approach for intrusion detection using recurrent neural networks." IEEE Access vol.5, pp.21954-21961, February 2017.

[9] Zheng, Zhuoyuan, YunpengCai, and Ye Li. "Oversampling method for imbalanced classification." Computing and Informatics, vol.34, no..5, pp. 1017-1037, 2016.

[10] Yuxin, Ding, and Zhu Siyi. "Malware detection based on deep learning algorithm." Neural Computing and Applications vol. 31, no.2, pp.461-472,February 2019.

[11] David, Omid E., and Nathan S. Netanyahu. "Deepsign: Deep learning for Automatic Malware Signature Generation and Classification." International Joint Conference on Neural Networks. IEEE, July 2015.

[12] Li, Yuancheng, Rong Ma, and Runhai Jiao. "A hybrid malicious code detection method based on deep learning." International Journal of Security and Its Applications vol.9, no.5, pp. 205-216, 2015.

[13] De Paola, Alessandra, et al. "Malware Detection through Low-level Features and Stacked DenoisingAutoencoders." Italian Conference on Cyber Security, ITASEC 2018, vol. 2058. CEUR-WS, February 2018.

[14] Patel, Ahmed, et al. "An intrusion detection and prevention system in cloud computing: A systematic review." Journal of Network and Computer Applications, vol.36 no.1, pp. 25-41, January 2013.

[15] Fiore, Ugo, et al. "Network anomaly detection with the restricted Boltzmann machine." Neurocomputing vol.122, pp. 13-23, 2013.

[16] Osada G., Omote K., Nishide T. "Network Intrusion Detection Based on Semi-supervised Variational Auto-Encoder", Foley S., Gollmann D., Snekkenes E. (eds) Computer Security – ESORICS 2017. ESORICS 2017. Springer, Cham Lecture Notes in Computer Science, vol 10493, August 2017.

[17] Zhang, Baoan, Yanhua Yu, and Jie Li. "Network Intrusion Detection Based on Stacked Sparse Autoencoder and Binary Tree Ensemble Method." 2018 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, May 2018.

[18] Jolliffe, Ian T., and Jorge Cadima. "Principal component analysis: a review and recent developments." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences vol. 374, no.2065, April 2016.

[19] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." 2015 Military Communications and Information Systems Conference (MilCIS). IEEE, pp. 1-6, November 2015.

AUTHOR'S PROFILE

Mr. S. SREENIVASA CHAKRAVARTHI, works as Assistant Professor in Center for Intelligent Computing, Sree Vidyanikethan Research Center, Sree Vidyanikethan Engineering College, Tirupati. He was awarded with B.E from Visveswaraiah Technological University, Belaguam, and M.Tech from Jawaharlal Nehru Technological University, Anantapur, Anantapuram. He is Part-Time Research Scholar in Vellore Institute of Technology, Chennai. He holds more than 16+ years of industry and academic experience. His research interest includes, Design of Algorithms, Machine Learning, Computational Intelligence, Cloud Computing, Network Security, Internet of Things, and Software Quality and Testing.

Dr. R. JAGADEESH KANNAN is senior professor in Vellore Institute of Technology, Chennai with 17+ years of teaching and industrial experience in the reputed organizations. He is Director - Innovation and Entrepreneurship Development Centre, and Chair - Computational Intelligence Research Group in VIT Chennai. He was an active player having a Liaison with various Institutions, R&D organizations and Industries, to promote various technical activities and an efficient Team Leader with outstanding organizational and excellent interpersonal skills to conduct any events such as seminars, symposiums, conferences, workshops, training programs, etc.

Towards a Classification View of Personalized e-Learning with Social Collaboration Support

Amal Al-Abri[1], Yassine Jamoussi[2], Zuhoor AlKhanjari[3], Naoufel Kraiem[4]
Department of Computer Science, Sultan Qaboos University
Muscat, Oman

***Abstract*—With the emergence of Web 2.0 technologies, interaction and collaboration support in the educational field have been augmented. These types of support embrace researchers to enrich the e-learning environment with personalized characteristics with the utilization of the collaboration support outputs. Achieving this requires understanding the existing environments and highlights their eminence. As a result, there are many attempts to state the current status of personalized e-learning environment from different perspectives. However, these attempts targeted a specific view and direction which failed to provide us with the general view of the adoption of personalized e-learning environment with the support of social collaboration tools. This paper provides a classified view of the current status of personalized e-learning environments which incorporate social collaboration tools for providing the personalization feature. The classification adopts four different views to carry out the classification; these views are subject, purpose, method, and tool. The findings show that the utilization of the user-generated contents and social interaction functionalities for personalization is tight and not fully consumed. In short, the potential of providing personalized learning with social interaction and collaboration features remains not fully explored.**

Keywords—Classification review; collaboration; personalized e-learning; social media

I. Introduction

Personalized e-learning environment is becoming more demanding in today's academic world [1]–[3]. According to Heller and his colleagues [4], the aim of personalized learning is to "tailor teaching to individual needs, interests, and aptitude." This form of online learning has the potential to serve the learners by providing a learning-teaching process according to the learner's needs as a medium of adaptation techniques to ensure the most effective knowledge transfer for each learner [5]. Besides, personalizable courseware requires connection with various tools like learning network for collaboration and performing learning tasks/activities. Therefore, by incorporating interactive Web 2.0 technologies like social media, personalized e-learning derived new opportunities in learning with the incorporation of collaborative learning activities [1].

With the emerging technology of web services like web 2.0 tools and enormous adoption of social collaboration tools for learning [6], the possibility of developing an e-learning platform for description, discovery, interaction, collaboration and interoperability of distributed, heterogeneous applications as services has been amplified enormously [7]–[10].

Collaborative learning is one of the learning styles motivated by the emergent of social media tools. According to [11], collaborative learning could be defined as "a variety of educational practices in which interactions among peers constitute the most important factor in learning, although without excluding other factors, such as the learning material and interactions with teachers". The characteristics of social media tools enhanced the adoption of collaborative learning activities in the educational field via the learning environments [12]. The concept of collaborative learning environment requires establishing a networked environment to facilitate active participation, interaction, and collaboration [13]. Besides, it also supports sharing and accessing learning resources among users [14]. The active participation opens the door for students to express themselves and share information related to their knowledge, preferences, and needs either explicitly or implicitly. This information is important to understand the characteristics of the learner. Consequently, provide personalization feature as it is highly demanded in today's educational environments [1], [15]–[17]. Thus, information related to the knowledge on the discussed topic and expressed opinion via like/dislike or textual expression needs to be extracted from social media.

From the scientific researcher's point of view, data-driven approaches to effectively facilitate personalization has only recently begun to emerge within higher education, especially with the integration of social collaboration tools [3]. Creating a personalized path especially with the involvement of social collaboration tools and services increased the amount of data and resources to be filtered and tailed to the learner's needs, which is a costly and complex task [3], [18], [19]. Especially since such involvement requires dealing with unstructured and noisy data generated during the collaboration. Therefore, understanding the integration of social collaboration tools in e-learning and the utilization of generated output during the use of these tools towards personalized e-learning is required. This type of investigation helps to understand the current status and adoption of these concepts to guide any further enhancement and development of new systems.

Despite the attempts carried out by different researchers to discuss the use of social collaboration in personalized e-learning, their discussions were limited to a specific dimension. For example, they focused on communication and collaboration techniques or user interface mechanisms or analysis techniques. As a result, provide a partial view of the integration which is still insufficient to have a full understanding of this concept. Therefore, this paper is attempting to give more advance view of the utilization of social collaboration in the personalized e-learning environment.

The paper is structured as follow: Section 2 discusses the literature review. Section 3 presents the approach proposed to classify the personalized e-learning systems with social collaboration supports systems. Section 4 provides a comparison of seven systems. In Section 5, a detailed discussion of the findings is presented. The paper is concluded in Section 6.

II. Literature Review

There have been many attempts by researchers in this field to state the current status of personalized e-learning environment from different perspectives. For example, In [20], the researchers focused on giving an overview of the use of adaptivity in e-learning by investigating whether it has been used and the extent to which it has been used. The researchers concluded that the majority of e-learning systems are not providing adaptivity features. Besides, the limited number of supported ones are not suitable for common e-learning scenarios as they are missing some standard features.

Another researcher like [21] discussed the state of art of personalization, particularly in the learning management system (LMS). This discussion focused on the limited personalization features provided by LMS like user interface and conditional activities feature which is controlled by the teacher to support the learning path.

A systematic review conducted by [22] to investigate the academic achievement in online higher education environments in accordance with the self-regulation (personalization) strategies followed by learners. This review considered some parameters like strategies, academic outcomes, participants, method design, course type & duration. However, this comprehensive review tackled peer learning as part of collaborative learning only partially.

The researchers in [23] conducted a survey on the employment of artificial intelligence techniques and related topics for adaptive e-learning (personalization). The discussion focused on one view of collaboration in the form of asynchronous or synchronous learning environments. Therefore, the survey is limited in the covered dimension.

As discussed in this section, the above-mentioned attempts, targeted a specific view and direction of personalization which provides a limited view of the personalized e-learning practices especially with the support of social collaboration tools. Therefore, the aim of this paper is to provide a comprehensive view of the current status of personalized e-learning environments which incorporate social collaboration tools for providing the personalization feature. This view will tackle the concept from four different angels as it is going to be discussed in Section 3.

III. Proposed Classification View

To investigate the utilization of social collaboration in the field of personalized e-learning, it is more appropriate to view it from different dimensions. This way the whole picture of the utilization will be clear and more understandable. For example, in a particular personalized e-learning system, it is helpful to know the target of the developed system, the reason behind the development, the manner where social collaboration has been integrated and the tools have been used for the collaboration. This can be achieved by classifying the view into different angels to cover the full picture.

The classification of social collaboration supports in the field of personalized e-learning using the web 2.0 functionalities can be derived using the classification framework as in [8]. The framework provides a multidimensional view to classifying personalized e-learning with social collaboration support systems. The view applies a faceted classification approach which can handle a multi-view and flexibility features [24] in analyzing the domain content of the targeted systems. It, as a result, enables the updating of the facet classification by adding new terms, modifying existing ones or even deleting unwanted ones without affecting other facets.

Faceted classification is defining the instantiated attribute classes with different terms. The facets are considered as perspectives, viewpoints, or dimensions of a particular domain. Each facet is measured by a set of relevant attributes. According to [25], “these attributes have values that are defined within a domain, whereby a domain may be a predefined type such as integral or Boolean, an enumerated type ({x, y, z}), or a structured type (Set {x, y})”. The set type allows characterizing the attributes using several values whose elements might belong to the enumerated type. Thus, a given collaboration approach might be positioned within a specific facet with two pairs of (attribute; value).

A. Classification Framework

The classification in the proposed framework is based on four different views as shown in Fig. 1. Each view captures a particular and relevant aspect of the systems. The four views are What (subject), Why (purpose), How (method), and Which (tool). Each view in the framework consists of a number of facets which present a set of attributes, and the attributes are defined by suitable values. Besides, there is an interconnection between the views which enforce the link between the different views in the framework.

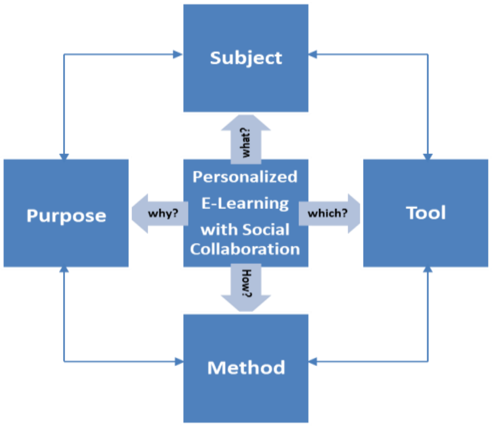


Fig. 1. Classification Framework for Personalized e-learning with Collaboration Support.

B. *Derived Facets from the Existing Framework*

Deriving from the classification framework proposed in our previous publication [8], the selected facets for the classification are depicted in Table I. The facets which represent the social collaboration characteristics in the four views as discussed in [8] give an overview of the attributes describing the different views of the classification framework. For instance, the subject view represents the "what" aspect of the framework by two facets related to the personalized e-learning with social collaboration support. These facets are actor facet and adaptability facet. The former consists of four attributes representing the different actors who can play an important role in social collaboration tools. It reflects the personalization aspect of the e-learning system either as a receiver or provider of the learning process. The latter facet represents the level of adaptability as a reflection of the personalization parameters provided by the tool.

The purpose view of the framework deals with the reason behind the development of the tool. In other words, it attempts to answer why the tool has been developed. From the social collaboration perspectives, the services facet is proposed to describe the purpose of the existing tool.

TABLE. I. SELECTED FACETS FROM CLASSIFICATION FRAMEWORK PRESENTED IN [8]

Concept/view	**Facet**	**Attribute (facet representation)**
Subject View	Actor Facet	- Learner: SET (ENUM {Students, Employee}) - Teacher: SET (ENUM {Instructor, Facilitator}) - Administrator: SET (ENUM {Institute, Developers}) - Experts: SET (ENUM {Researchers, Experts})
	Adaptability Facet	- Adaptability level: SET:{High, Medium, Low}
Purpose View	Services Facet	- Monitoring: BOOLEAN - Learner tracking: BOOLEAN - Visualization: SET (ENUM{Graphical, Algorithm}) - Grading & Evaluation: SET (ENUM{Scores, Analysis})
Method View	Collaboration Methods Facet	- Interaction: SET (ENUM{Collaborative writing, Communicating chatting and social interaction, file sharing, brainstorming, sharing links and bookmarks, media sharing, computer-intensive e-learning services}) - Integration: SET (ENUM{Plugins, Stand-alone, Mashups})
Tool View	(LOs) Facet	- Granularity: SET {aggregation level, complexity} - Authoring: SET {content (multimedia objects, real world objects), metadata} - Standards: SET {SCORM, ISM-LD, IEEE LOM, SOA} - Language: SET {language()}

The method view describes the "how" angle of the framework. It presents the methods adapted for the social collaboration which provide support to delegate the personalization feature to the learners. Collaboration methods fact derived from the existing framework represents two attributes (interaction and integration) as methods for delivering the social collaboration feature.

The tool view in the classification framework elucidates the "which" part of the framework. This indicates the importance of the tools used by the actors based on the method applied to achieve the purpose of the personalized e-learning system. The extracted LOs facet represents the resources as a tool in the social collaboration process to support the delivery of personalized learning resources.

The complete list of the derived facets for each view and their respected attributes are depicted in Table I.

As the vital concept in the classification view presented in this paper is the personalization, there is a need to look closely into the personalization aspect from different angles related to the support derived through the social collaboration tools. Besides, the previous classification focuses on the collaboration concept rather than the personalization aspect. Therefore, advanced facets related to personalized e-learning are generated based on the four views of the classification framework as per the discussion in the next sub-section.

C. *Generated Facets*

In the personalized learning environment (PLE), the Learning place is an aggregation of communication and collaboration tools, shared resources, services, and people [26]. According to Becta, (2007), PLE should also help to satisfy individual needs (personalization). Therefore, when assessing any e-learning environment developed for personalization purpose, there is a need to verify the availability of the following parameters.

- Communication and collaboration: Promoting communication and collaboration are one of the functionalities in e-learning which may enhance the process to provide personalized e-learning environment [26], [27]. Generated data during collaboration can be a valuable source of information to understand the learner’s characteristics and concepts under discussion, which will in turn help in providing personalization feature. Referring to the classification framework, this functionality belongs to the method view of the systems. Specifically, the collaboration method facet.
- Resources: Learning resources or objects are the key source of knowledge to be shared between learners during the learning process [28]. Therefore, learning resources considered as an important facet in an e-learning environment. For personalized e-learning, learning resources are the package to be delivered to the learners to avoid heterogeneous part of the available resources [29].
- Services: The e-learning environment should facilitate the interaction and management of the learning tool through the availability of vital services, like web

application for interactions [26], tools for monitoring, tracking learners' progress, visualizing and representing data, analyzing and evaluating the obtained data [10], [27]. The service facet in the purpose view of the classification framework is discussed.

- People: Learners and teachers are the main parties in the learning process. However, when integrating collaboration support, other parties can also be involved in the process, allowing for more engagements, which will eventually help to enrich the discussions. People or actors like friends and experts as discussed in the classification framework.

- Adaptation: adaptation techniques and parameters enlighten the path to follow and the aspects of learners to take under considerations when providing personalization [30]. In the area of collaboration support, the user-generated content is utilized to extract information related to the parameters used for personalization as well as discovering and representing the domain under discussion [7]. Measuring to what extent the system is adaptive to the learners' needs is very important as discussed in the classification framework.

Based on the identified parameters of personalized learning with collaboration support systems and the classification framework proposed to compare collaboration approaches in e-learning, five facets (actor, adaptability, service, collaboration method and LOs) from the four views will be considered. The mapping between the personalized learning environment parameters and the classification framework is depicted in Fig. 2. The dotted area shows the mapping part, displaying the views and the corresponding facets from the classification framework.

To understand the personalized e-learning environment with collaboration support based on the identified facets from the classification framework, the attributes and values for each facet need to be stated. For this comparison, all the attributes and values will be the same, except for the adaptability facet. As the main focus is the personalization aspect in these systems, some attributes need to be added to illustrate how personalization has been carried out in these systems apart from the adaptability level. The added attributes for this purpose are:

- *Learner's characteristic* which specifies the parameters considered in providing personalization. According to [31], the most popular and useful features which can distinguish the learner as an individual are; the learner's knowledge, interests, goals, background, preferences, and individual traits like (cognitive style and learning style).

- *Component* that represents the model in a semi/automatically generated during the personalization process. According to [32], the core of the architecture of an adaptive application is formed by three closely linked components: the domain model (DM); the user model (UM); and the adaptation model (AM). In a learning application, for instance, the user model will keep track of the user's knowledge of each of the concepts in the domain model. The adaptation model defined the rules that state how the adaptation must be performed and the actual adaptation performs by the adaptive engine [32].

- *Technique* for expressing the methods adopted to provide personalization. The adaptation technologies are adopted from three areas which are intelligent tutoring systems (ITS), adaptive hypermedia (AH), and adaptive collaboration support (ACLS) [33].

The added attributes in the adaptability facet can be classified as the followings:

Characteristic: SET (ENUM {knowledge, interests, goals, background, preferences, individual traits})

Component: SET (ENUM {domain model, user model, adaptation model})

Technique: SET (ENUM {intelligent tutoring systems, adaptive hypermedia, adaptive collaboration support})

The detailed view of the facets incorporated in this paper is depicted in Fig. 3. These identified facets are used to compare 7 systems as discussed in the following section (Section 3).

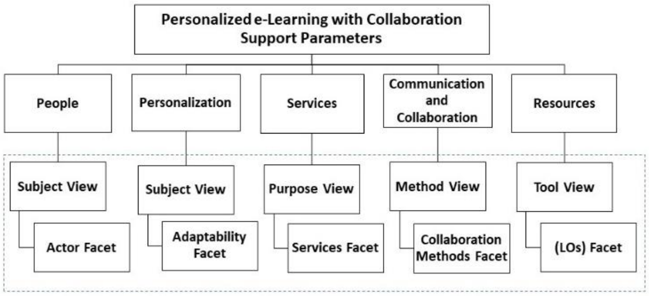


Fig. 2. Mapping the Personalized e-learning Parameters with the Classification Framework Views.

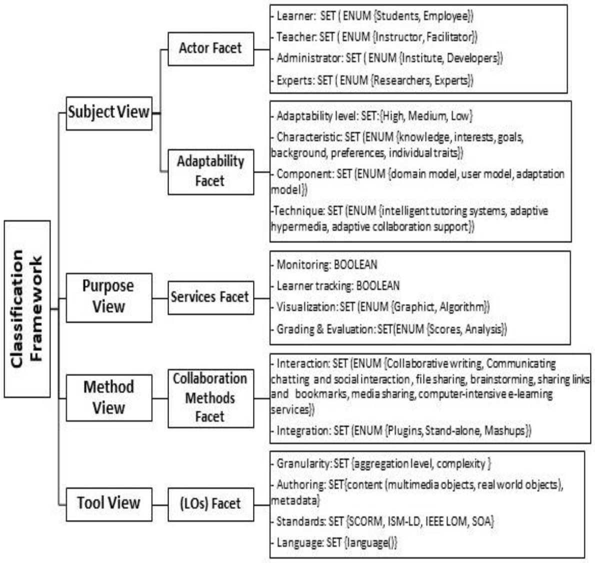

Classification Framework
Subject View
Actor Facet
- Learner: SET (ENUM {Students, Employee})
- Teacher: SET (ENUM {Instructor, Facilitator})
- Administrator: SET (ENUM {Institute, Developers})
- Experts: SET (ENUM {Researchers, Experts})
Adaptability Facet
- Adaptability level: SET:{High, Medium, Low}
- Characteristic: SET (ENUM {knowledge, interests, goals, background, preferences, individual traits})
- Component: SET (ENUM {domain model, user model, adaptation model})
-Technique: SET (ENUM {intelligent tutoring systems, adaptive hypermedia, adaptive collaboration support})
Purpose View
Services Facet
- Monitoring: BOOLEAN
- Learner tracking: BOOLEAN
- Visualization: SET (ENUM {Graphict, Algorithm})
- Grading & Evaluation: SET(ENUM {Scores, Analysis})
Method View
Collaboration Methods Facet
- Interaction: SET (ENUM {Collaborative writing, Communicating chatting and social interaction , file sharing, brainstorming, sharing links and bookmarks, media sharing, computer-intensive e-learning services})
- Integration: SET (ENUM {Plugins, Stand-alone, Mashups})
Tool View
(LOs) Facet
- Granularity: SET {aggregation level, complexity }
- Authoring: SET{content (multimedia objects, real world objects), metadata}
- Standards: SET {SCORM, ISM-LD, IEEE LOM, SOA}
- Language: SET {language()}

Fig. 3. Detailed Facets of the Classification Framework.

IV. Comparison of Personalized E-learning with Social Collaboration Support

As a reflection to the aim of the proposed classification framework discussed in this paper, seven systems have been selected to investigate the usefulness of the framework in giving the complete view on the targeted area of discussion. The selected systems are providing the personalization feature supported by social collaboration tools.

The systems selected for the comparison applying the proposed framework are presenting an attempt to add the social dimension by integrating web 2.0 functionalities with the adaptive/personalized learning techniques. The systems are WHURLE 2.0, SLAOS, GRAPPLE, Topolor, ALEF, SALT and Protus. The description of each of these systems and how social collaboration support is used for personalization are given below.

WHURLE 2.0 [34] consists of five independent Web services that collaborate with each other to tailor a unique view of the learning content for a given learner, and a delivery service (LMS) where the learner views this adaptive content. The framework incorporated Adaptive Educational Hypermedia Systems and web services (SOA). The WHURLE 2.0 has been tested for its adaptation and social collaborative interactive functionalities by providing it with the LMS's built-in tools such as a forum, chat, and wiki to perform the social activity. However, the aim of the study was only to investigate the extent to which students make use of the collaboration tools and if they aid in their learning process. Besides, the system separated the personalization and collaboration processes.

SLAOS [35] is a framework aiming to bring together three features which are; web 2.0, e-learning and adaptive personalization. The framework extended the adaptive hypermedia framework by integrating a social layer. This layer has features like collaborative authoring and social annotation. The authors' approach allows students to be part of the authoring stage but with some sets of privileges. The collaborative facilities in SLAOS rely on Web 2.0 techniques, such as group-based authoring, cooperation in creating the courses, tagging the content, and rating. However, the support provided on the domain modeling level and the ability to support collaboration based on user-generated content is limited.

GRAPPLE [36] is another framework that supports the learning process via (adaptive guidance and personalized content). The framework consists of two key components of which are GRAPPLE Adaptive Learning Engine (GALE), where the content adaptation is performed, and GRAPPLE User Modeling Framework (GUMF), in charge of managing user model data. LMS, GUMF, and GALE are communicated through GRAPPLE Event Bus (GEB). The framework aggregates and enriches the user modeling in GUMF by embodying Mypes service which exploits dataspaces to connect, aggregate, align and enrich user profile information from social media tools [37]. However, the focus was on the personal information located in the user profile which is not providing enough information about user knowledge and other characteristics.

Topolor [38], is a framework that introduces Web 2.0 tools into an adaptive educational hypermedia system. The framework is a layered based architecture consisting of two layers; storage layer and runtime layer. Topolor has a Facebook-like appearance and supports social annotation and collaborative learning by introducing the Affiliate Model. The framework provides a social e-learning environment, where learners can comment on a topic, ask or answer a question, create and share notes. It also supports learning content adaptation, learning path adaptation and peer adaptation. However, Topolor does not consider the use of data on preferred items for adaptation. Besides the framework considers the look of only one social interaction tool which is Facebook as a mean of a simple interaction between learners.

ALEF [39] is an Adaptive LEarning Framework. It is a framework for creating adaptive and highly interactive web-based learning systems. The system proposes a generic model, namely domain model, based on lightweight semantics which opens new possibilities of automated course metadata creation and student model. ALEF combines different learning activities (such as learning from explanatory texts, questions or exercises) along with the highly interactive and social environment of the Web 2.0. The framework provides personalized learning by recommending learning objects tailored to the student needs to be based on the student's knowledge. However, the framework suggested a limited number of social interaction mechanisms and it is not supporting the learning object authored by students.

SALT [40] is a framework for social learning which integrates social network functionality with traditional adaptive educational hypermedia to engage students into learning through teaching and adapt learning pathways to individual student needs based on collective learning experiences. The users (student and teacher) interact with each other by contributing to constructing a small learning content in the form of mini-lessons (lesslet). SALT implements self-organized personalization through learning pathways. However, the research focus is mainly on crowdsourcing and scalability issues like grouping students based on similar user's performance.

Protus (PRogramming TUtoring System) [41] is an intelligent web-based programming tutoring system. The system integrates collaborative tagging technique to provide personalized recommended learning resources. The tagging mechanism of the lessons adopted in the system provides information related to the learner's interest. This information is the key to identify the learning style. The system also uses a test to identify the knowledge level. Consequently, the learner's interest and previous knowledge level are used to provide personalized recommendations.

Table II summarizes the comparison between the above seven selected systems. The discussion on the finding is presented in the results and discussion section.

TABLE. II. COMPARATIVE TABLE OF THE SELECTED SYSTEMS USING THE CLASSIFICATION FRAMEWORK

System	**Actor**				**Personalization**				**Services**				**Collaboration method**		**Resource (LO)**			
	Learner	Teacher	Administrator	Expert	Characteristics	Components	Techniques	Level	Monitoring	Learner tracking	Visualization	Grading & Evaluation	Interaction	Integration	Granularity	Authoring	Standards	Language
WHURLE 2.0 (2009)	S	I	In	N	K	UM	AH	H	MN	N	N	Sc	Co	PI	A	Cn, Md	SOA	PHP XML XSLT
SLAOS (2011)	S	I	N	Ot	K	UM	AH	H	MN	LT	N	Sc An	Co	SA	A	Cn, Md	N	MySQL CGI (C++) RDF
GRAPPLE (2012)	S	I	D	N	K	UM	AH	H	N	N	G	Sc	N	PI	A	Cn, Md	SOA	XML XHTML RDF
Topolor (2013)	S	I	N	N	K P	UM	AH	H	MN	LT	G	Sc An	Co	SA	A	Cn Md	N	PHP SQL
ALEF (2014)	S	I	N	N	K GO	UM DM	AH	H	MN	LT	G	An	Co BK	SA	A	Cn Md	N	XML
SALT (2017)	S	I	N	N	K	UM	ITS AH	H	MN	N	G	Sc An	CW Co	SA	A	Cn Md	N	C# ASP.NET
Protus (2018)	S	I	N	N	K IT	UM	ITS	H	MN	N	N	An	BK	SA	A	Cn Md	N	Java

TABLE II: (CONTINUED)

Table Abbreviations:

Subject View

- S: Student
- E: Employee
- I: Instructor
- F: Facilitator
- In: Institute
- D: Developer
- R: Researcher
- Ot: Outsider
- K: Knowledge
- IR: Interests
- GO: Goals
- B: Background
- P: preferences
- IT: Individual traits
- DM: Domain model
- UM: User model
- AM: Adaptation model
- ITS: Intelligent tutoring systems
- AH: Adaptive hypermedia
- ACLS: Adaptive collaboration support
- H: High
- M: Medium
- L: Low
- N: Not Applicable

Purpose View

- MN: Monitoring
- LT: Learner tracking
- Vis: Visualization (G: Graphics, A: Algorithm)
- G&E: Grading & Evaluation (Sc: Scores, An: Analysis)

Method View

- CW: Collaborative writing
- Co: Communicate
- FS: File sharing
- BR: Brainstorming
- BK: Bookmarks
- MS: Media sharing
- EL: e-learning
- PI: Plugins
- SA: Stand-alone
- MA: Mashups

Tool View

- A: Aggregation level
- C: Complexity
- Cn: Content
- Md: Metadata
- SC: SCORM
- LD: ISM-LD
- LOM: IEEE LOM
- OS: OpenSocial

V. Results and Discussion

The attempt to incorporate the social collaboration concept to provide the personalization feature is growing up among researchers in the educational field. However, while designing the systems, not all the required aspects have been considered. As illustrated in Table II, the finding of the comparison of the selected seven systems with respect to the four views can be summarized as follows:

A. *Subject View*

1) The main roles are played by the student as a learner and instructor as a teacher in all compared systems.

2) Most systems do not provide information related to the responsible party for managing the system except for WHURLE 2.0 and GRAPPLE.

3) Only SLAOS gives permission to an outsider to access the developed system.

4) All systems deliver a high level of personalization.

5) Most systems provide personalization based on one parameter which is the knowledge level except three of them which are Topolor, ALEF, and Protus.

6) Collaboration support has been utilized to generate/update the user model in all compared systems except the lightweight domain modeling proposed in ALEF. This indicates the limited support for semi/automatic construction of the domain model.

7) Most of the compared systems (85%) use adaptive hypermedia (AH) technique for personalization except for Protus which uses Intelligent tutoring technique (ITS).

8) An overall consideration of the facets (Actor & Personalization) in subject view by compared systems is shown in Fig. 4. The figure illustrates the focus on the main actors of an e-learning environment which are the learner and teacher (41% each) and less attention to expert (6%) only. It also shows the high attention played by all systems in delivering personalization considering the four attributes representing the personalization facet.

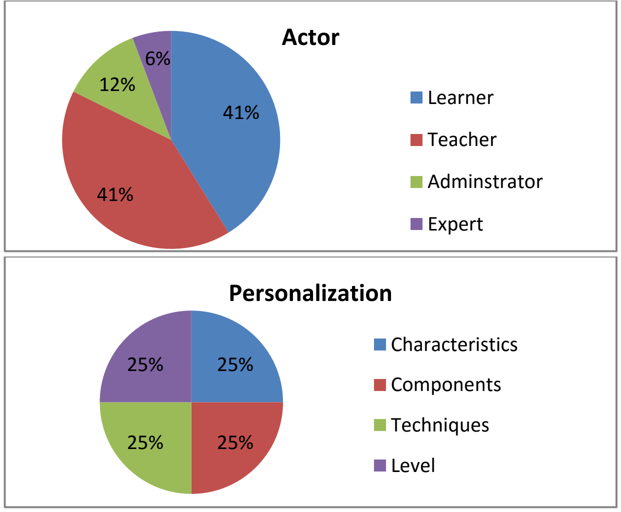


Fig. 4. Consideration of the Facets (Actor and Personalization) in the Subject View.

B. *Purpose View*

1) Almost 85% of the compared systems targeted monitoring (MN) purpose and 40% targeted learner tracking (LT) purpose.

2) The visualization service provided by 57% of the compared systems was based on graphics visualization.

3) Both scores and analysis parameters were considered with respect to grading and evaluation service by compared systems either together as in SLAOS, Topolor, and SALT or one of the parameters.

4) The overall consideration of service fact is quite promising as depicted in Fig. 5. However, learning tracking attribute needs to be deliberated more to add more value to the developed system as it is currently getting the least attention (15%) among other attributes in this fact.

C. *Method View*

1) Around 70% of the systems incorporated communication technique as an interaction method.

2) The integration of the collaboration method using a stand-alone technique is quite promising for data analysis in most systems especially by implementing semi-structured discussion as in Topolor, ALEF, and Protus. However, the data analysis by utilizing the user-generated content was limited.

3) As depicted in Fig. 6, the overall consideration of the Collaboration method facet in the method view is acceptable as both attributes almost equally considered (46% interaction & 54% integration). However, there is a limited consideration in the values of the different attributes as shown in the comparison table (Table II).

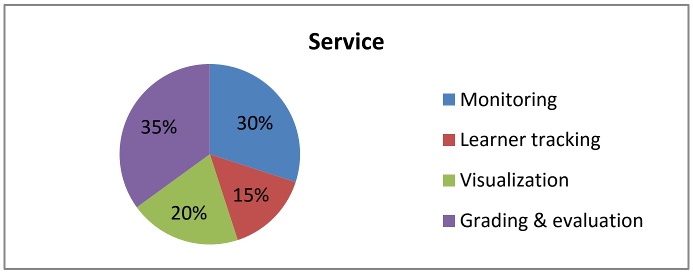


Fig. 5. Consideration of the Service Facet in the Purpose View.

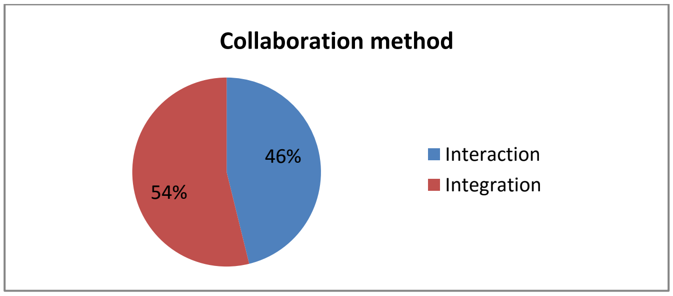


Fig. 6. Consideration of the Collaboration Method Facet in the Method View.

D. Tool View

1) The aggregation has been considered for granularity by all compared systems.

2) Content as a mean of the multimedia object has been considered for granularity by all compared systems.

3) With respect to the standard used for development, no information has been provided by researchers except for WHURLE 2.0 and GRAPPLE; they adopt SOA.

4) Various languages have been used for the implementation of the compared systems depend on the suitability to the functionalities and the embedded environment.

5) The overall consideration of the different attributes of the resource (LO) facet as shown in Fig. 7 is quite good. However, there is a need to focus on adapting the learning resource development standards like SCORM, IMS-LD, IEE LOM and SOA [12] when developing the systems as it is presenting only 9% consideration.

The overall remarks concluded from the classification are quite inspiring. In short, there is an acceptable level of social interaction proposed by the studied systems in terms of the interaction methods and the tools used. However, utilization of the user-generated content and social interaction functionalities for personalization is tight and not fully consumed. For instance, it is used to update one parameter of personalization and only one researcher used it for domain model construction. In fact, the potential of providing personalized learning based on social interaction and collaboration features remains not fully explored. The main issues which are required to be addressed in future systems based on the findings are:

1) Provide personalization based on more than one parameter, for example, the knowledge level and individual traits (Learning style).

2) The parameters (Knowledge level and learning style) could be obtained from user-generated content during collaboration.

3) Collaboration support can be the source to generate/update the user model and the construction of the domain model.

4) Provide a semi/structure collaboration tool to facilitate the ongoing discussions as in Topolor, ALEF, and Protus.

5) The integration of LMS is advisable to provide personalized e-learning environment within the commonly used environment for learning as incorporated by WHURLE 2.0 and GRAPPLE.

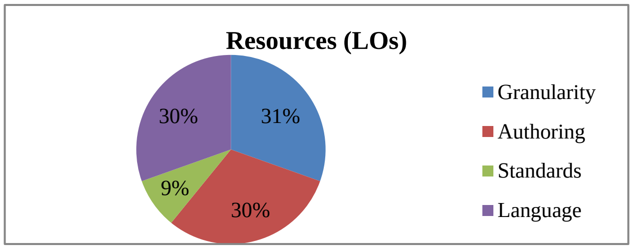


Fig. 7. Consideration of the Resource (LOs) Facet in the Tool View.

VI. Conclusion and Future Work

Applying the social dimension with adaptive learning by integrating Web 2.0 functionalities is the core concept to be studied in the era of the digital age. Therefore, understanding the adoption of personalized e-learning environment with the support of social collaboration tools is required for further development and upgrading. Despite the attempts carried out by different researchers to review the current status of personalized e-learning environments, they are targeting only a specific domain. Consequently, lacking provides a general or detailed view of the adoption of collaboration tools in this field. This paper gives a classified view of the current status of personalized e-learning environments which incorporates social collaboration tools for providing the personalization feature. The view classified the studied systems using four views: subject, purpose, method, and tool which gives a more comprehensive overview of the systems' functionalities. The comparison of the seven selected systems shows that the adoption of social interaction and collaboration tools is quite good in the educational field. It also shows that the utilization of the user-generated contents and social interaction functionalities for personalization is tight and not fully consumed. In general, the potential of providing personalized learning with social interaction and collaboration features remains not fully explored. The results indicate the importance of utilizing these remarks to achieve more advanced personalized e-learning systems.

Therefore, tackling the unconsumed functionalities like the utilization of user-generated contents during social interaction for personalization purpose is tagged as our future work. Besides, the flexible proposed framework can be expanded to cover more technical dimensions like the techniques and algorithms used for analyzing the social collaboration context towards personalization.

Acknowledgment

This work is supported by the Ministry of Manpower, Sultanate of Oman.

References

[1] B. Phillips, "Beyond Classroom Learning: Personalized Learning Through Digital Technologies," 2016.

[2] M. Kravcik, O. C. Santos, and J. G. Boticario, "4th International Workshop on Personalization Approaches in Learning Environments (PALE 2014)," 2014.

[3] L. Johnson, S. Adams Becker, M. Cummins, V. Estrada, A. Freeman, and C. Hall, "NMC Horizon Report: 2016 Higher Education Edition," Austin, Texas New Media Consort., 2016.

[4] J. Heller, C. Steiner, C. Hockemeyer, and A. Dietrich, "Competence-based knowledge structures for personalised learning," Int. J. ELearning, vol. 5, no. 1, p. 75, 2006.

[5] F. Mödritscher and F. Wild, "Personalized e-learning through environment design and collaborative activities," in Symposium of the Austrian HCI and Usability Engineering Group, 2008, pp. 377–390.

[6] Y. Jamoussi, Z. Al-khanjari, and N. Kraiem, "A Framework to Evaluate E-learning Based on Social Networking," vol. 2, no. 4, pp. 26–42, 2014.

[7] A. Al-Abri, Z. Al-Khanjari, N. Kraiem, and Y. Jamoussi, "A scheme for extracting information from collaborative social interaction tools for personalized educational environments," in Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017, 2017, vol. 2017-Janua.

[8] A. Al-Abri, Y. Jamoussi, N. Kraiem, and Z. Al-Khanjari, “Comprehensive classification of collaboration approaches in E-learning,” Telemat. Informatics, vol. 34, no. 6, 2017.

[9] Z. Al-Khanjari, Y. Al-Roshdi, and N. Kraiem, “Virt-res: Developing extended architectural design for computer science virtual resources using SOA,” Int. J. Softw. Eng. Its Appl., vol. 8, no. 9, pp. 125–136, 2014.

[10] E. Popescu, “Providing collaborative learning support with social media in an integrated environment,” World Wide Web, vol. 17, no. 2, pp. 199–212, 2014.

[11] P. Dillenbourg, S. Järvelä, and F. Fischer, “The evolution of research on computer-supported collaborative learning,” in Technology-enhanced learning, Springer, 2009, pp. 3–19.

[12] A. Al-Abri, Z. Al-Khanjari, Y. Jamoussi, and N. Kraiem, “Developing a Collaborative Learning Environment Using Web Services Techniques.,” JSW, vol. 11, no. 9, pp. 870–882, 2016.

[13] Z. Du, X. Fu, C. Zhao, Q. Liu, and T. Liu, “Interactive and collaborative e-learning platform with integrated social software and learning management system,” in Proceedings of the 2012 International Conference on Information Technology and Software Engineering, 2013, vol. 212, pp. 11–18.

[14] M. Masud, “Collaborative e-learning systems using semantic data interoperability,” Comput. Human Behav., vol. 61, pp. 127–135, 2016.

[15] A. Al-Abri, Z. Al-Khanjari, Y. Jamoussi, and N. Kraiem, “Identifying Learning Styles from Chat Conversation using Ontology-Based Dynamic Bayesian Network Model,” in 2018 8th International Conference on Computer Science and Information Technology (CSIT), 2018, pp. 1–8.

[16] Z. A. Al-Khanjari, “Developing a common personalization framework for the e-application software systems,” J. Emerg. Technol. Web Intell., vol. 5, no. 2, pp. 188–195, 2013.

[17] N. Dabbagh and A. Kitsantas, “Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning,” Internet High. Educ., vol. 15, no. 1, pp. 3–8, 2012.

[18] A. Kahaei, “Design of Personalization of Massive Open Online Courses,” 2014.

[19] J. Heller, B. Mayer, and D. Albert, “Competence-based knowledge structures for personalised learning,” in 1st International ELeGI Conference on Advanced Technology for Enhanced Learning, 2005, p. 8.

[20] D. Hauger and M. Köck, “State of the Art of Adaptivity in E-Learning Platforms.,” in LWA, 2007, pp. 355–360.

[21] C. Limongelli, F. Sciarrone, and G. Vaste, “Personalized e-learning in moodle: The moodle_LS system,” J. E-Learning Knowl. Soc., vol. 7, no. 1, pp. 49–58, 2011.

[22] J. Broadbent and W. L. Poon, “Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review,” Internet High. Educ., vol. 27, pp. 1–13, 2015.

[23] K. Colchester, H. Hagras, D. Alghazzawi, and G. Aldabbagh, “A survey of artificial intelligence techniques employed for adaptive educational systems within e-learning platforms,” J. Artif. Intell. Soft Comput. Res., vol. 7, no. 1, pp. 47–64, 2017.

[24] W. Denton, “How to make a faceted classification and put it on the web,” Online (November 2003) http//www. miskatonic. org/library/facet-web-howto. html, 2003.

[25] O. Saidani, R. S. Kaabi, N. Kraiem, and Y. Baghdadi, “A multidimensional framework to classify goal-oriented approaches for services,” in Computer Applications Technology (ICCAT), 2013 International Conference on, 2013, pp. 1–5.

[26] D. Gillet, “Personal learning environments as enablers for connectivist MOOCs,” in Information Technology Based Higher Education and Training (ITHET), 2013 International Conference on, 2013, pp. 1–5.

[27] Becta, “Learning platforms and personalising learning. An essential guide.” p. 6, 2007.

[28] Z. A. Al-Khanjari and N. S. Kutti, “Re-Engineering Learning Objects for Re-Purposing,” Int. J. Eng. Res. Appl., vol. 4, no. 8, pp. 103–111, 2014.

[29] K. Anandakumar, K. Rathipriya, and A. Bharathi, “A survey on methodologies for personalized e-Learning recommender systems,” Int. J. Innov. Res. Comput. Commun. Eng., vol. 2, no. 6, 2014.

[30] E. Kurilovas, S. Kubilinskiene, and V. Dagiene, “Web 3.0–Based personalisation of learning objects in virtual learning environments,” Comput. Human Behav., vol. 30, pp. 654–662, 2014.

[31] P. Brusilovsky and E. Millán, “User models for adaptive hypermedia and adaptive educational systems,” in The adaptive web, Springer, 2007, pp. 3–53.

[32] P. De Bra and J.-P. Ruiter, “Aha! adaptive hypermedia for all,” Proc. AACE WebNet Conf., no. DECEMBER 2001, pp. 262–268, 2001.

[33] P. Brusilovsky, “Adaptive educational systems on the world-wide-web: A review of available technologies,” in Proceedings of Workshop" WWW-Based Tutoring" at 4th International Conference on Intelligent Tutoring Systems (ITS’98), San Antonio, TX, 1998.

[34] M. Meccawy, “A service-orientated architecture for adaptive and collaborative e-learning systems .,” University of Nottingham, 2009.

[35] A. I. Cristea and F. Ghali, “Towards Adaptation in E-Learning 2 . 0,” 2010.

[36] P. De Bra et al., “GRAPPLE Learning Management Systems Meet Adaptive Learning Environments.pdf,” Intell. Adapt. Educ. Syst., pp. 133–160, 2013.

[37] F. Abel, N. Henze, E. Herder, G.-J. Houben, D. Krause, and E. Leonardi, “Building blocks for user modeling with data from the social web,” CEUR Workshop Proc., vol. 609, 2010.

[38] L. Shi, A. I. Cristea, J. G. K. Foss, D. Al Qudah, and A. Qaffas, “A social personalized adaptive e-learning environment: a case study in topolor,” vol. 7641, pp. 1–17, 2013.

[39] M. Bielikovà et al., “ALEF: From application to platform for adaptive collaborative learning,” Recomm. Syst. Technol. Enhanc. Learn. Res. Trends Appl., no. 1, pp. 195–225, 2014.

[40] E. Karataev and V. Zadorozhny, “on Crowdsourcing,” vol. 10, no. 2, pp. 128–139, 2017.

[41] A. Klašnja-Milićević, M. Ivanović, B. Vesin, and Z. Budimac, “Enhancing e-learning systems with personalized recommendation based on collaborative tagging techniques,” Appl. Intell., vol. 48, no. 6, pp. 1519–1535, 2018.

An Efficient Design of RPL Objective Function for Routing in Internet of Things using Fuzzy Logic

Adeeb Saaidah[1], Omar Almomani[2]
Computer Network and Information Systems Department
Faculty of Information Technology, The World Islamic
Sciences & Education University Amman, Jordan

Laila Al-Qaisi[3]
Computer Science Department, Faculty of Information
Technology, The World Islamic Sciences & Education
University Amman, Jordan

Mohammed Kamel MADI[4]
Faculty of Computing and Information Technology, Sohar University, Sohar, Sultanate of Oman

***Abstract*—The nature of the Low power and lossy networks (LLNs) requires having efficient protocols capable of handling the resource constraints. LLNs consist of networks that connect different type of devices which has constraints resources such as energy, memory and battery life. Using the standard routing protocols such as Open Shortest Path First (OSPF) is inefficient for LLNs due to the constraints that LLNs need. So, IPv6 Routing Protocol for Low-Power and Lossy Networks (RPL) was developed to accommodate these constraints. RPL is a distance vector protocol that used the object functions (OF) to define the best tree path. However, choosing a single metric for the OF found to be unable to accommodate applications requirements. In this paper, an enhanced (OF) is proposed namely; OFRRT-FUZZY relying on several metrics combined using Fuzzy Logic. In order to overcome the limitations of using a single metric, the proposed OFRRT-FUZZY considers node and link metrics. Namely, Received Signal Strength Indicator (RSSI), Remaining Energy (RE) and Throughput (TH). The proposed OFRRT-FUZZY is implemented under Cooja simulator and then results were compared with OF0, MHROF in order to find which OF provides more satisfactory results. And simulation results show that OFRRT-FUZZY outperformed OF0 and MHROF.**

***Keywords*—RPL; Objective Function; IOT; fuzzy logic; LLNs**

I. Introduction

The revolution of networks is changing rapidly, starting from Internet to smart phone until Internet of Things (IoT) glare. This revolution is driven by the capabilities of (IoT) to provide benefits for any system either that user or business [1]. This benefits to ease human life through allowing devices to make tasks instead of human [2]. IoT represents the future of computing and communications in several fields from wireless sensors to nanotechnology. IoT comes from the two words “Internet” and “Things” which means that IoT networks connect different things which forms a heterogonous system such as sensors, monitors and smart devices [3][1]. These heterogonous devices can be able to collects data, make communication, computation and ultimate decision making [4]. There are many definitions for IoT, the best one is explained by [5]: “An open and comprehensive network of intelligent objects that have the capacity to auto-organize, share information, data and resources, reacting and acting in face of situations and changes in the environment” [1]. As a result, the rapid increase of number of things which are connected to the network while some of them have limited resources, needs more attention in order to handle all these constraints for LLNs.

One of the critical issues in IoT is the constraints for those devices that connected in LLNs [6][7]. These constraints such as battery power which means it has limited life time, short transmission range, in addition to noisy environment arrangement in LLNs, needs particular protocol to handle the increasing of number of IoT devices and the constraints in order to provide a robust routing of data and efficiency [5][7][8][9]. As a result, the Internet Engineering Task Force (IETF) developed a routing standardized protocol for 6LoWPAN networks which is called RPL [10]. RPL overcomes and handle all the constraint requirements that need for LLNs. Furthermore, It can adapt the variation of link and node metrics over time. Considering that LLNs nodes can have one or several parent which forms a route to the root.

RPL is the main candidate for acting as the standard routing protocol for IPv6 based LLNs, like wireless sensor networks [11]. It has been proposed as a tree routing protocol by the roll working group. It is also an extensible and flexible single path protocol. Nevertheless, it only saves an optimal routing at a specific time. It chooses a parent node of the best parent nodes [12]. It uses an OF to find the best path. Moreover, the chosen metrics to define the best path aren’t defined by the working group. RPL uses a set of constraints and metrics through specific OF for building a Destination Oriented Directed Acyclic Graph (DODAG). The OFs choose the best parent of nodes for building and optimizing the route. However, the standard OF suffer from several limitations which are attributed to the single metric usage. In order to solve the limitations of employing a single metric, a new OF based on combined metrics using Fuzzy Logic is proposed. This OF, considers the node and the link metrics, RSSI, RE and TH.

The second part of this study aims at discussing RPL routing protocol. The third part aims at reviewing the relevant works, whereas the fourth part aims at discussing the metrics. The fifth part sheds a light on the proposed OFRRT_FUZZY algorithm. The sixth part presents the outcomes of the developed simulation. The seventh part presents the conclusion.

II. RPL Routing Protocol

RPL was designed by ROLL working group from IETF to cover the limited resources that attached to the connected devices in LLNs called RPL [13][14]. These limited resources, such as battery life, memory and connectivity to the Internet. The existing routing protocol such as OSPF unable to handle these limitations for these reasons RPL considered as a standard routing protocol for LLNs. RPL is an IPV6 distance vector routing protocol [15] that use the mechanism of Directed Acyclic Graphs (DAGs) [16][17] to apply a tree structure between nodes. Hence each node could be having more than one parent each one could be a next hop in a path to reach a sink. The nodes organized as (DODAGs) [18] where the best parent are chosen based on the metrics. The information exchange between nodes by using the following ICMPv6 control messages [10][19].

- DODAG Information Object (DIO): It's employed for storing the information needed for creating the upward routes of DODAGs such as current Rank, DODAG ID, etc.
- Destination Advertisement Object (DAO): It's employed for sending destination information (path information) upwards to the root along the DODAG.
- DODAG Information Solicitation (DIS): It's employed for enabling the node to solicit a DODAG information object (DIO) from a reachable neighbor.
- Destination Advertisement Object Acknowledgement (DAO-ACK): Unicast message sent by recipient node to acknowledge DAO message.

Fig. 1 show the node working once receives DIO messages and to select the best parents based on calculated rank.

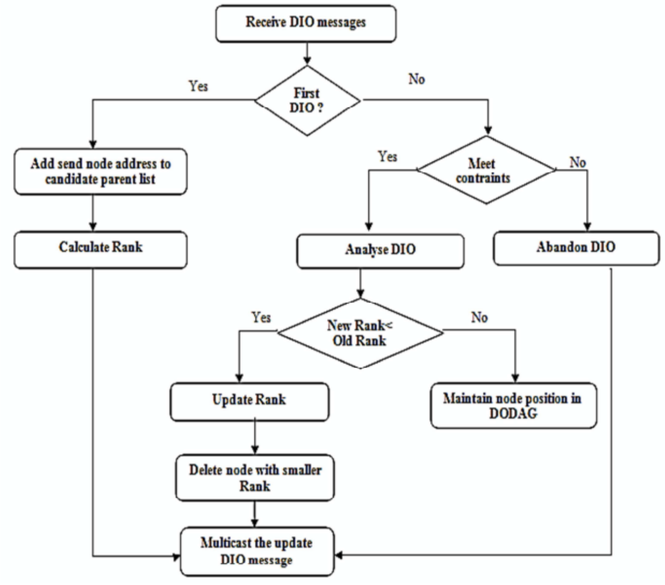

Receive DIO messages
Yes
First DIO ?
No
Add send node address to candidate parent list
Yes
Meet contraints
No
Calculate Rank
Analyse DIO
Abandon DIO
Yes
New Rank< Old Rank
No
Update Rank
Maintain node position in DODAG
Delete node with smaller Rank
Multicast the update DIO message

Fig. 1. Different Operations Performed after Receiving a DIO and the Process of Rank Calculation in RPL.

III. Related Work

OF is considered as vital part of RPL. It determines major decisions such as parent selection and forwarding path. OF is a hot topic researchers in the field got interested to investigate in order to improve RPL performance. Moreover, fuzzy logic was chosen to be a major addition to OF in RPL.

In [20] Gaddour et al., designed a Fuzzy Logic OF which combined a set of metrics namely; Hop Count, End-to-End Delay, Node Energy and Link Quality. These were used as fuzzy parameters to configure a routing decision. The OF designed, took the application requirements in to consideration for the purpose of finding best path to destination. The evaluation was conducted using large-scale test bed in Contiki OS and Cooja Simulations showed that OF had registered a great improvement in the RPL-based LLNs compared to other used OF.

While [21] Kamgueu et al., proposed a fuzzy inference system for getting better performance than other used systems. They saw that routing systems usually tend to rely on network lifetime without referring to other network performance metrics. As a result, they combined expected transmission delay, count, and node's remaining power. Their Implementation was conducted on Contiki OS and simulations were performed on Cooja. Results showed the combined metrics OF registered significant improvements among one metric OF especially the ETX scenario.

In [22] Kamgueu et al., developed a new RPL OF which uses a fuzzy inference system for combining several metrics (i.e. the expected transmission delay, count, and node's remaining power). The assessment was conducted through using a real sensor network which was deployed in an indoor environment. The results showed that combined fuzzy OF performed better than ETX based routing on energy efficiency, packet loss ratio, routing stability, and end-to-end delay.

In [23] Lamaazi and Benamar, proposed a combined metrics based on fuzzy logic OF. It was designed with node and link metrics (i.e. hop count and energy consumption & expected transmission count). Their results showed that combined OF-Fuzzy outperformed OF based ETX and OF based energy consumption in terms of overhead and Packet Delivery Ratio (PDR). It was found that OF-Fuzzy participated in equalizing nodes' energy consumption through the network.

Another research conducted by [24] Aljarrah proposed a multi fuzzy logic model for OF for RPL (Ml-FL) consisting of three vital metrics: node-oriented metrics, channel-oriented metrics and link-oriented metrics for uncasing. The Ml-FL chose the best parent for uncast through nine individual metrics. Three other parameters were used to define each of the nine metrics to ensure effective parent node selection. To overcome fuzzy logic complexity, multiple fuzzy logic blocks were processed in parallel. Moreover, an enhanced- BMRF algorithm is proposed with the minimum of delay and duplicate packets. IEEE 802.15.4 standard was applied over OMNeT++ simulator for assessing the effectiveness of the proposed RPL. The outcomes reached through the proposed RPL are promising in terms of energy, end to end delay, hop count, packet delivery ratio and packet loss rate.

In [25] Fabian et al., stated that amount of data represented by IoT is rapidly growing. While wireless sensors discompose a great challenge due to their diverge radio conditions, limited energy and computational capabilities. These were reasons for proposing fuzzy logic OF for dynamic adaptation of wireless network changing environment. Their results showed that fuzzy OF improved performance comparing to others in terms of throughput by 15% and 14% of (PDR) not mentioning energy consumption.

Another combined Fuzzy Logic OF was proposed by [11] Lamaazi and Benamar for overcoming the limitation of the standard OF which relies on single metric. The node and link metrics (i.e. Energy Consumption, Hop Count and Expected Transmission Count) were used in building the proposed OF-EC. Effectiveness of newly proposed OF-EC was shown through simulations compared to MRHOF, ENTOT, and OF-FUZZY. Major improvements were found in RPL performance in terms of PDR, convergence time, network lifetime, overhead, latency and energy consumption. It was found that OF-EC retained the efficiency of RPL totally independent from transmission rate & network topology. Furthermore, OF-EC performed better than other proposals in terms of energy consumption among all nodes through network.

In [5] Sankar and Srinivasan, proposed an energy aware fuzzy logic RPL called (FLEA-RPL). It took several routing metrics into consideration; expected transmission count (ETX), residual energy (RER), and load, for choosing the best route. Fuzzy logic was applied over selected metrics to choose best route for efficient data transmission through network. Experiments were conducted using COOJA simulator for assessing proposed FLEA-RPL. Then set of comparisons with other similar protocols: MRHOF (ETX) based RPL (MRHOF-RPL) and FL-RPL. Results showed 10-12% improvement in: network lifetime and 2-5% in packet delivery ratio for FLEA-RPL.

IV. Choosing Metrics

The metrics were chosen in order to get better performance measures. Using node metrics only cannot be able to accommodate the LLN constraints. As a result, combining link and node metrics is applied in this research paper. As previously stated, the proposed OFRRT-FUZZY considers the link and node metrics, namely, (RSSI), (RE) and (TH).

The major challenging requirements of RPL protocol are mobility support, reliable routing, energy-efficient routing and achieve higher throughput [24]. Usually sensors networks have several constraints; a critical one is that sensor nodes use batteries. Second, those sensors maybe deployed both unattended and in large numbers. That makes it difficult neither to change nor to recharge batteries in sensors. As a result, all processes, systems and communication protocols of sensor networks or sensors must take into consideration the minimizing power consumption. In addition, RSSI gives an indication of the power level that is received by the antenna. This means the higher RSSI level, the radio signal shall be stronger and thus, the destination shall be closer. (RSSI) can also be used as a measurement for wireless link quality [7]. Sriniv et al. [26] presents facts about RSSI in estimating link quality. RSSI is considered available during a packet reception with neither any impact on energy consumption, or additional hardware, nor throughput. In addition, this metric looks intuitive: stronger is the received signal, closer is the transmitter and weaker is the received signal, further is the transmitter. RSSI is employed in several standards for identifying the time on which the amount of radio energy in the channel is considered below a specific threshold at which point the node is clear to send.

Next section discusses the proposed OFRRT_FUZZY algorithm.

V. The Proposed Ofrrt-Fuzzy Algorithm

The OFRRT_FUZZY was developed based on fuzzy logic to enhance the performance results of the standard OFs (OF0 AND MRHOF) and choose the optimal path to reach the destination. Standard OFs usually depend on using only a single metric. the single metric approach doesn't meet all the application requirements [27][28]. Based on the node metric, OF0 is capable of providing a loosy quality of the link. While with MRHOF, the nodes shall provide a better link quality. However, it may be associated with a higher level of energy consumption. Due to these reasons, a combination of node and link metrics are proposed [23]. In particular, the proposed OFRRT-FUUZY uses fuzzy inference process(FIP), which is by definition "a process of mapping from a given input to an output, using the theory of fuzzy sets" [29]. Accordingly, the proposed method uses three input linguistic variables, namely, (RE), (TH) and (RSSI) to calculate a single output linguistic variable (Best path). The Mamdani-style FIP is applied to achieve this objective. It is one of the most commonly used fuzzy inference techniques. It includes four steps: crisp input fuzzification, rule evaluation, rule output aggregation, and defuzzification as shown in Fig. 2 the pseudo code of proposed Algorithm are discussed in Table I.

A. *Fuzzification*

In this step, crisp inputs, which are TH, RSSI and RE, are processed to determine the degree to which these inputs belong to each appropriate fuzzy set. Every linguistic variable has it is own Universe of Discourse (UOD), which determines its range of values. The value of a fuzzy set is provided based on the behavior of a linguistic variable. The fuzzy sets for the inputs and output are shown as follows:

TH={LOW, MID, HIGH}

RSSI={CONNECTED,TRANSITIONING,DISCONNECTED
}

RE= {SMALL, AVERAGE, HIGH}

Each linguistic variable has it is own UOD that clarifies its boundaries. The membership functions of the input linguistic variables (TH, RSSI and RE) are presented in Fig. 3, Fig. 4 and Fig. 5, respectively and the output linguistic variable (best path) are presented in Fig. 6. For computational simplicity, linguistic variables are frequently represented by triangles or trapezoids. A trapezoid is used in the proposed OFRRT-FUZZY algorithm. The UOD for the TH input linguistic variable ranges from 0 to 1500 according to Maximum Transmission Unit (MTU). On the other side, the UOD for the

RSSI input linguistic variable ranges from -50 to -100, according to [30]. Last metric is RE the value is range between 0 to 250 as in [20]. The boundaries of the fuzzy sets and membership functions are chosen by domain experts [29].

B. Fuzzy Rule Evaluation

In this phase, the fuzzified inputs are applied to the rules based on knowledge, which is created by domain experts. In general, when the number of inputs increases in fuzzy logic, the complexity of rule application also increases. To solve this problem, multiple input single output is used [31]. To obtain a satisfactory rule base in fuzzy logic, rules should generally exhibit the following properties. First, rules should be complete, which indicates that rules must cover all system behavior. Second, rules should be consistent, which indicates that all rules must be logically valid. Each rule consists of two parts, the antecedent part of fuzzy rules, which is represented by "if" body rule, and consequent part (output). The antecedent part includes the input variables. Thus, the fuzzified inputs, together with their membership degrees, are applied to the first part to obtain the degree of membership in the consequent part. Some of rules are presented in Table II.

.

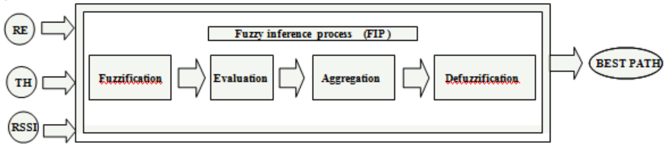
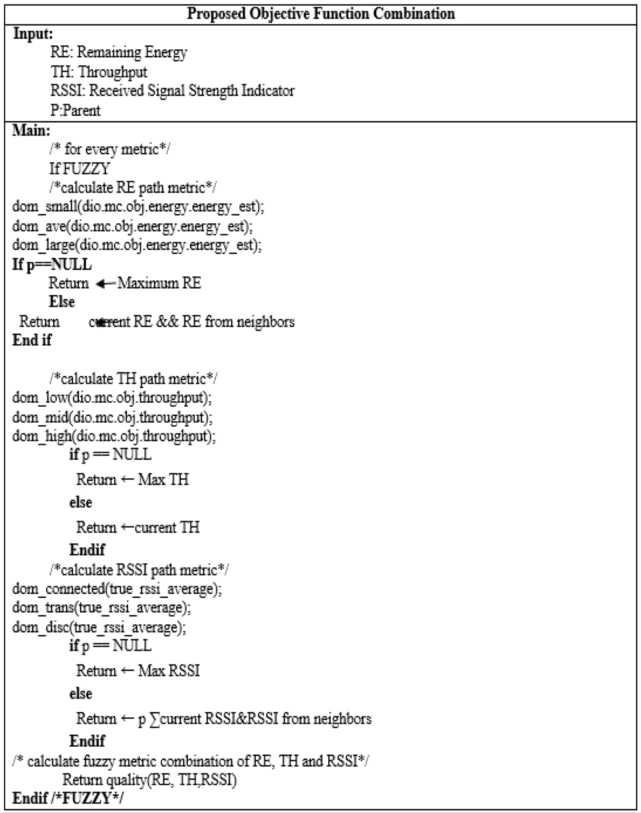


Fig. 2. Fuzzy Inference Process.

TABLE. I. PSEUDO CODE OF PROPOSED ALGORITHM



```
Proposed Objective Function Combination
Input:
        RE: Remaining Energy
        TH: Throughput
        RSSI: Received Signal Strength Indicator
        P:Parent
Main:
        /* for every metric*/
        If FUZZY
        /*calculate RE path metric*/
dom_small(dio.mc.obj.energy.energy_est);
dom_ave(dio.mc.obj.energy.energy_est);
dom_large(dio.mc.obj.energy.energy_est);
If p==NULL
        Return ←Maximum RE
        Else
  Return      current RE && RE from neighbors
End if

        /*calculate TH path metric*/
dom_low(dio.mc.obj.throughput);
dom_mid(dio.mc.obj.throughput);
dom_high(dio.mc.obj.throughput);
            if p == NULL
             Return ← Max TH
            else
             Return ←current TH
            Endif
        /*calculate RSSI path metric*/
dom_connected(true_rssi_average);
dom_trans(true_rssi_average);
dom_disc(true_rssi_average);
            if p == NULL
             Return ← Max RSSI
            else
             Return ← p ∑current RSSI&RSSI from neighbors
            Endif
/* calculate fuzzy metric combination of RE, TH and RSSI*/
           Return quality(RE, TH,RSSI)
Endif /*FUZZY*/
```

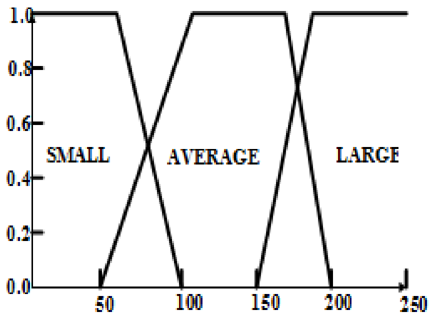


Fig. 3. Memberships Function of RE.

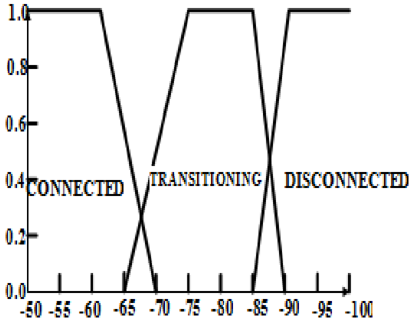


Fig. 4. Memberships Function of RSSL

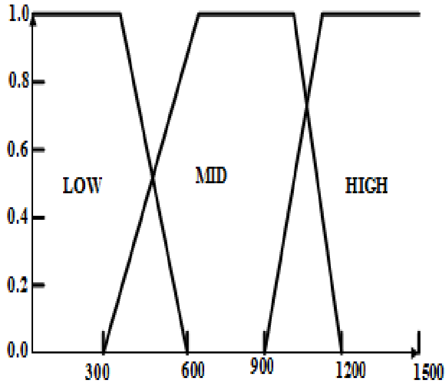


Fig. 5. Memberships Function of TH.

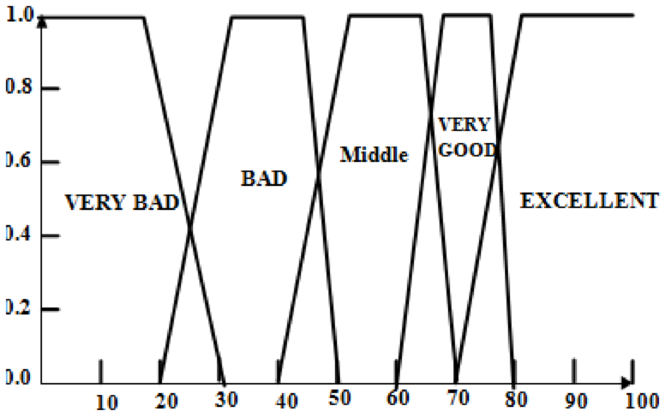


Fig. 6. Memberships Function of Best Path.

TABLE. II. SOME OF FUZZY RULES FOR OFRRT-FUZZY

Condition	Result
IF RSSI is CONNECTED and this LOW and RE is SMALL	Bad
IF RSSI is CONNECTED and this MID and RE is AVERAGE	Very Good
IF RSSI is CONNECTED and this HIGH and RE is LARGE	Excellent
IF RSSI is TRANSITIONIN and this LOW and RE is SMALL	Bad
IF RSSI is TRANSITIONIN and this LOW and RE is AVERAGE	Middle
IF RSSI is TRANSITIONIN and this LOW and RE is LARGE	Middle
IF RSSI is DISCONNECTED and this LOW and RE is SMALL	Very bad
IF RSSI is DISCONNECTED and this LOW and RE is AVERAGE	Bad
IF RSSI is DISCONNECTED and this LOW and RE is LARGE	Bad

To obtain a good degree of output, the minimum operation will be adopted because the operation applied between the linguistic inputs is the “AND” operation for all the rules.

C. Rule Output Aggregation

Previously, a degree of membership is assigned to each consequent rule, and thus, combining all the outputs membership values from all the rules into a single fuzzy set is required. This process combines the inputs list of membership values with the single output fuzzy set for each output variable.

D. Defuzzification

The final step in FIP is to select a defuzzification method. The input for this step is the output from the previous step, which is a fuzzy set for every output linguistic variable, whereas the output from this step is a crisp value for each output linguistic variable. The defuzzfication method derives the crisp value to numerical value which is representing the fuzzy value of the linguistic output variable. In the OFRRT-FUZZY proposed algorithm, the center of gravity (COG) is selected to find the numerical value of output. COG is the most popular approach because it finds the point at which a vertical line will divide the aggregate set into two masses that are equal [29]. COG is expressed in the following formula [29]:

$$\mathrm{COG} = \frac{\sum_a^b \mu_A(\mathrm{Si}) \times \mathrm{Si}}{\sum_a^b \mu_A(\mathrm{Si})} \quad (1)$$

where, µA(S) refers to the membership function of elements Si in sub-set A and S represent the degrees of membership.

Next section illustrate the performance evaluation of OFRRT_FUZZY

VI. PERFORMANCE EVALUATION

A. Simulation Setting

A set of simulation experiments were designed for examining the performance level of the proposed OFRRT-FUZZY in RPL protocol for IoT and compare it with the existing OFs such as OF0 and MRHOF. This simulation is carried out by COOJA simulation under Contiki OS. Contiki OS is an open source emulator that was designed for the IoT technology. Table III present the values for all parameters employed through the simulation experiments.

TABLE. III. COOJA SIMULATION PARAMETERS

20 Network simulator	21 COOJA under Contiki OS (2.7)
22 Objective function	23 OF0 , MRHOF(ETX) and Proposed OF
24 Number of nodes	25 15,30,45
26 Nodes type	27 Sky mote
28 Topology	29 Grid
30 Radio medium (Wireless Channel)	31 Unit Disk Graph Medium (UDGM) with distance lose
32 TX Ratio	33 100%
34 RX Ratio	35 100%
36 Transmission Range	37 50 meter
38 Interference Range	39 100 m
40 Simulation Times	41 Minutes

B. Chosen Performance Metrics

For evaluating the RPL performance level, the metrics must be chosen accurately to show the pros and cons of each RPL OFs. Therefore, the four metrics listed below were selected to evaluate The OFRRT_FUUZY, OF0 and MRHOF.

1) Average latency: It refers to the average time a transmitted packet consumes from sender node to sink node. The following equation can be used for calculating it:

$$\text{Average latency} = \frac{\sum_{k=1}^{n} \text{recv time(k)} - \text{send time(k)}}{\text{total packet received}} \quad (2)$$

2) Packet delivery ratio (PDR): It refers to the ratio of nodes’ number of received and sent packets. The following equation can be used for calculating it:

$$\text{PDR} = \frac{\text{total received packets at the sink}}{\text{Total sent packets from senders}} \quad (3)$$

3) Energy consumption (mJ): Energy anode requires to exchanges data through the network between nodes. The following equation can be used for calculating it.

$$\text{Energy Consumption} = (\text{Transmit} * 19.5\text{mA} + \text{Listen} * 21.5\text{mA} + \text{CPU\_time} * 1.8\text{mA} + \text{LPM} * 0.0545\text{mA}) * 3\text{V}/(32768) \quad (4)$$

4) Control traffic overhead: Represents the total number of control messages DIO, DAO &DIS used by ICMPv6, it is calculated based on the following equation.

$$\text{Control Traffic Overhead} = \sum_1^n DIO + \sum_1^n DIS + \sum_1^n DAO \quad (5)$$

C. Results and Discussion

1) Average latency: The first evaluation metric for this study is to exam the impact of average latency for OFRRT-FUZZY, OF0 and MRHOF with different size of network. Fig. 7 illustrates the measurement of average latency for each network size with different OF (OFRRY-FUZZY, OF0 and MRHOF). The result shows that the OFRRT-FUZZY has lower average latency as compare to OF0 and MRHOF for each network size and MRHOF has higher average latency. In MRHOF case Select path from sender to sink based on ETX may get long time thus increase average latency. In case of OF0 select path from sender to sink based on number of hops will not decrease latency too much because some of nodes may get congested. In case of OFRRT-Fuzzy Select path from sender to sink based on combination of three matrices using fuzzy logic, less control traffic messages generated thus leads to less average latency.

2) Packet Delivery Ratio (PDR): Here the impact of PDR is exam for OFRRT-fuzzy, OF0 and MRHOF to observe the network reliability. The network size is varying. Fig. 8 Illustrates measure the PDR for each network size of each OF (OFRRT-Fuzzy, OF0 and MRHOF).

The result shows that OFRRT-Fuzzy has higher PDR as compare to OF0 and MRHOF for each network size and OF0 has low PDR. In OFRRT-Fuzzy allows for less packets loss due to use more than one metrics for selecting best path and this lead to higher PDR, MRHOF has more packet loss compare to OFRRT-Fuzzy due to continues calculation of best path. OF0 has the lower PDR, because it is produce more packets loss due to congestion that might occurs in the sender specially the senders that far from sink.

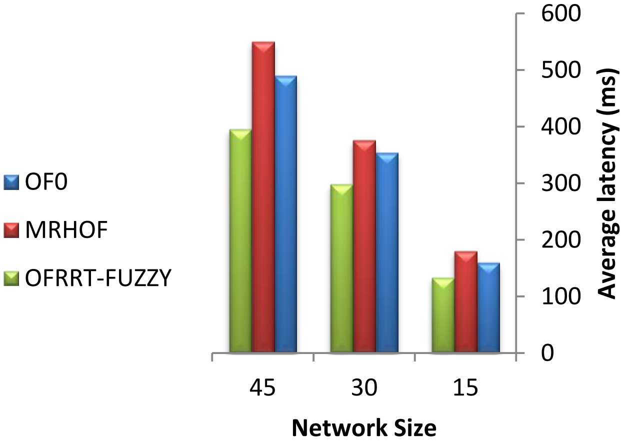


Fig. 7. Average Latency for OF0, MRHOF and OF –FUZZY.

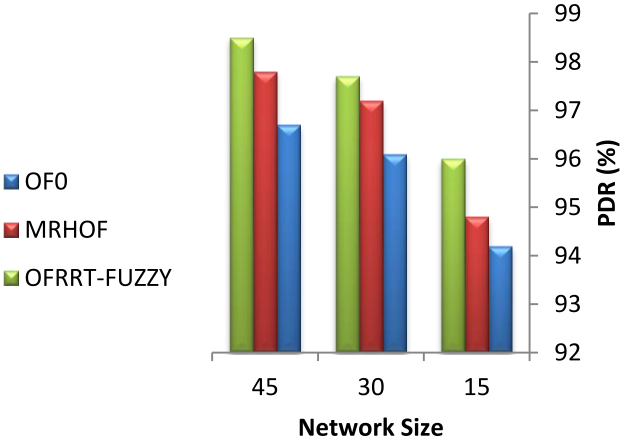


Fig. 8. Packet Delivery Ratio (PDR) For OF0, MRHOF and OF –FUZZY.

3) Energy Consumption (mJ): Here the impact of average power consumption is exam for OFRRT-FUZZY, OF0 and MRHOF. The network size is varying. The average power consumption calculating by considering the average power consumption used in each node for each network size of every OF (OFRRT-Fuzzy, OF0 and MRHOF).

Fig. 9 illustrates average power consumption for OFRRT-Fuzzy, OF0 and MRHOF of different network size, the result shows that the OFRRT-Fuzzy has lower average power consumption as compare to OF0 and MRHOF for each network size and MRHOF has higher average power consumption and with each network size and it is almost same as OF0. In OFRRT-Fuzzy allows less control message due using of combination metrics to select path from sender to skin therefore less CPU_ Power used relating to the node processing. HRHOF has higher average power consumption because the ETX calculation needed more process than hop and this lead to more average power consumption.

4) Control traffic overhead: In this section exam the impact of control traffic overhead for OFRRT-fuzzy, OF0 and MRHOF with different network size. The control traffic overhead is calculating by the sum of DIO, DIS and DAO for each network size of every OF (OFRRT-Fuzzy,OF0 and MRHOF).

Fig. 10 illustrates control traffic overhead for OFRRT-Fuzzy, OF0 and MRHOF of different network size, the result shows that OFRRT-Fuzzy has lower control traffic overhead as compare to OF0 and MRHOF for each network size and MRHOF has higher control traffic overhead. RPL produce more control traffic messages during DODAG set up, Once DODAG is constructed, less ICMPv6 are produce. In OFRRT-Fuzzy Calculating combination of three metrics using fuzzy logic for select path from sender to skin needs less control messages as compared to MRHOF and OF0 because DODAG are constructed faster. MRHOF needs more control messages to complete DODAG constructed as compare to OF-FUZZY and OF0.

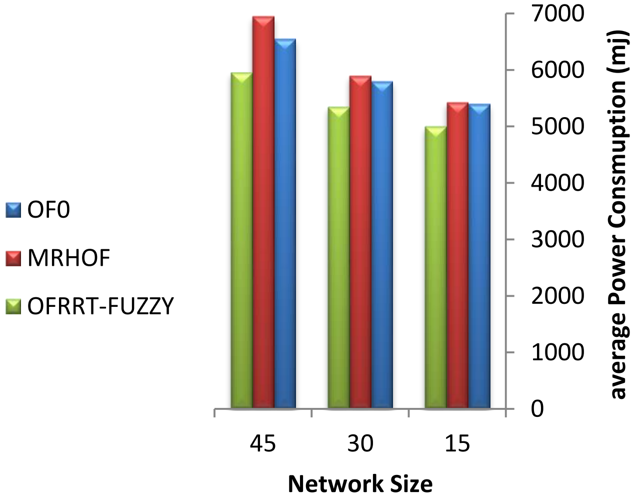


Fig. 9. Average Power Consumption for OF0, MRHOF and OF –FUZZY.

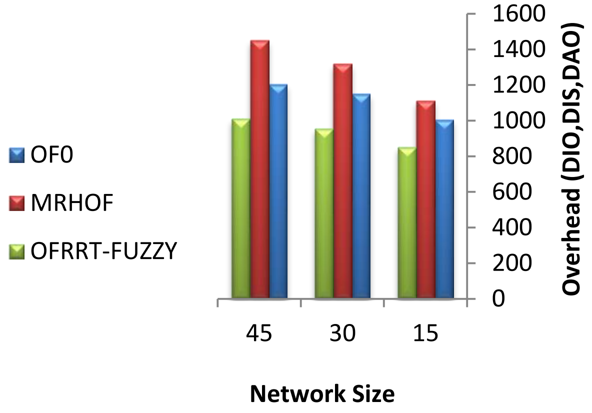


Fig. 10. Overhead for OF0, MRHOF and OF –FUZZY.

VII. Conclusion

LLNs require certain type of protocols to work with. That is due to the fact of resource constraints. Mainly LLNs includes various types of devices which has different constraints as they are functioning. Standard routing protocol does not meet such needs. That is exactly why RPL was introduced. It is a distance vector protocol that finds the best path through OF. Many researchers are referred to in the related work section considered setting an OF with a single metric, that is either node or link. This paper proposed an OFRRT-FUZZY based on fuzzy logic which combines both link and nodes metrics namely; RE, RSSI and TH. The fuzzy logic method -which relies on a fuzzy membership which determines a set of rules for the combination-was adopted. Set of simulation experiments were conducted on Cooja and the results showed that OFRRT-fuzzy provide more satisfactory performance than OF0 and MHROF in terms of latency, PDR, overhead & Power consumption.

References

[1] S. Madakam, R. Ramaswamy, and S. Tripathi, “Internet of Things (IoT): A literature review,” J. Comput. Commun., vol. 3, no. 05, p. 164, 2015.

[2] S. Agrawal and M. L. Das, “Internet of Things—A paradigm shift of future Internet applications,” in 2011 Nirma University International Conference on Engineering, 2011, pp. 1–7.

[3] D. Airehrour, J. Gutierrez, and S. K. Ray, “Secure routing for internet of things: A survey,” J. Netw. Comput. Appl., vol. 66, pp. 198–213, 2016.

[4] D. Airehrour, J. A. Gutierrez, and S. K. Ray, “SecTrust-RPL: A secure trust-aware RPL routing protocol for Internet of Things,” Futur. Gener. Comput. Syst., vol. 93, pp. 860–876, 2019.

[5] S. Sankar and P. Srinivasan, “Fuzzy Logic Based Energy Aware Routing Protocol for Internet of Things,” Int. J. Intell. Syst. Appl., vol. 10, no. 10, p. 11, 2018.

[6] M. D\'\iaz, C. Mart\'\in, and B. Rubio, “State-of-the-art, challenges, and open issues in the integration of Internet of things and cloud computing,” J. Netw. Comput. Appl., vol. 67, pp. 99–117, 2016.

[7] S. Sankar and P. Srinivasan, “Composite Metric Based Energy Efficient Routing Protocol for Internet of Things,” Int. J. Intell. Eng. Syst., vol. 10, no. 5, pp. 278–286, 2017.

[8] H. Lamaazi, N. Benamar, A. J. Jara, L. Ladid, and D. El Ouadghiri, “Challenges of the internet of things: IPv6 and network management,” in 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2014, pp. 328–333.

[9] J. H. Kong, L.-M. Ang, and K. P. Seng, “A comprehensive survey of modern symmetric cryptographic solutions for resource constrained environments,” J. Netw. Comput. Appl., vol. 49, pp. 15–50, 2015.

[10] H. Lamaazi, N. Benamar, M. I. Imaduddin, and A. J. Jara, “Performance assessment of the routing protocol for low power and lossy networks,” in 2015 International Conference on Wireless Networks and Mobile Communications (WINCOM), 2015, pp. 1–8.

[11] H. Lamaazi and N. Benamar, “OF-EC: A novel energy consumption aware objective function for RPL based on fuzzy logic.,” J. Netw. Comput. Appl., 2018.

[12] Z. Wang, L. Zhang, Z. Zheng, and J. Wang, “Energy balancing RPL protocol with multipath for wireless sensor networks,” Peer-to-Peer Netw. Appl., vol. 11, no. 5, pp. 1085–1100, 2018.

[13] G. G. Lorente, B. Lemmens, M. Carlier, A. Braeken, and K. Steenhaut, “BMRF: Bidirectional multicast RPL forwarding,” Ad Hoc Networks, vol. 54, pp. 69–84, 2017.

[14] P. Di Marco, G. Athanasiou, P.-V. Mekikis, and C. Fischione, “MAC-aware routing metrics for the internet of things,” Comput. Commun., vol. 74, pp. 77–86, 2016.

[15] M. B. Yassein, S. Aljawarneh, and others, “A new elastic trickle timer algorithm for Internet of Things,” J. Netw. Comput. Appl., vol. 89, pp. 38–47, 2017.

[16] M. Zhao, P. H. J. Chong, and H. C. B. Chan, “An energy-efficient and cluster-parent based RPL with power-level refinement for low-power and lossy networks,” Comput. Commun., vol. 104, pp. 17–33, 2017.

[17] S. Tennina, O. Gaddour, A. Koubâa, F. Royo, M. Alves, and M. Abid, “Z-Monitor: A protocol analyzer for IEEE 802.15. 4-based low-power wireless networks,” Comput. Networks, vol. 95, pp. 77–96, 2016.

[18] A. Oliveira and T. Vazão, “Low-power and lossy networks under mobility: A survey,” Comput. networks, vol. 107, pp. 339–352, 2016.

[19] H. Fotouhi, D. Moreira, M. Alves, and P. M. Yomsi, “mRPL+: A mobility management framework in RPL/6LoWPAN,” Comput. Commun., vol. 104, pp. 34–54, 2017.

[20] O. Gaddour, A. Koubâa, N. Baccour, and M. Abid, “OF-FL: QoS-aware fuzzy logic objective function for the RPL routing protocol,” in 2014 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2014, pp. 365–372.

[21] P. O. Kamgueu, E. Nataf, T. Djotio, and O. Festor, “Fuzzy-based routing metrics combination for RPL,” 2014.

[22] P.-O. Kamgueu, E. Nataf, and T. N. Djotio, “On design and deployment of fuzzy-based metric for routing in low-power and lossy networks,” in 2015 IEEE 40th Local Computer Networks Conference Workshops (LCN Workshops), 2015, pp. 789–795.

[23] 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), 2017, pp. 1459–1464.

[24] E. Aljarrah, “Deployment of multi-fuzzy model based routing in RPL to support efficient IoT,” Int. J. Commun. Networks Inf. Secur., vol. 9, no. 3, pp. 457–465, 2017.

[25] P. Fabian, A. Rachedi, C. Gueguen, and S. Lohier, “Fuzzy-Based Objective Function for Routing Protocol in the Internet of Things,” in 2018 IEEE Global Communications Conference (GLOBECOM), 2018, pp. 1–6.

[26] K. Srinivasan and P. Levis, “RSSI is under appreciated,” in Proceedings of the third workshop on embedded networked sensors (EmNets), 2006, vol. 2006.

[27] P. O. Kamgueu, E. Nataf, T. D. Ndié, and O. Festor, “Energy-based routing metric for RPL,” INRIA, 2013.

[28] J. P. Yunis and D. Dujovne, “Energy efficient routing performance evaluation for LLNs using combined metrics,” in 2014 IEEE Biennial Congress of Argentina (ARGENCON), 2014, pp. 741–746.

[29] N. Michael, Artificial Intelligence A Guide to Intelligent Systems, vol. 321204662. 2005.

[30] I. H. Urama, H. Fotouhi, and M. M. Abdellatif, “Optimizing RPL objective function for mobile low-power wireless networks,” in 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017, vol. 2, pp. 678–683.

[31] M.-S. Kim and S.-G. Kong, “Parallel-structure fuzzy system for time series prediction,” Int. J. Fuzzy Syst., vol. 3, no. 1, pp. 331–340, 2001.

Modelling the Enterprise Architecture Implementation in the Public Sector using HOT-Fit Framework

Hasimi Sallehudin[1], Nurhizam Safie Mohd Satar[2]
Faculty of Technology and Science Technology
Universiti Kebangsaan Malaysia, Selangor, Malaysia

Farashazillah Yahya[5]
Faculty of Computing and Informatics
Universiti Malaysia Sabah, Malaysia

Nur Azaliah Abu Bakar[3], Rogis Baker[4]
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia[3]
Faculty of Defence Studies and Management
National Defence University of Malaysia
Kuala Lumpur, Malaysia[4]

Ahmad Firdause Md Fadzil[6]
Faculty of Economy and Management Science
University of Sultan Zainal Abidin
Kuala Terengganu
Terengganu, Malaysia

***Abstract*—Enterprise architecture is very important to the public sector's IT systems that are developed, organized, scaled up, maintained and strategized. Despite an extensive literature, the research of enterprise architecture is still at the early stage in public the sector and the reason to explain the acceptance, as well as the understanding of the implementation level of EA services still remains unclear. Therefore, this study examines the implementation of EA by measuring the Malaysian public sector's influence factors of EA. Grounded by the Human-Organization-Technology (HOT-Fit) Model, this study proposes a conceptual framework by decomposing Human characteristics, Organizational characteristics and Technological characteristics as main categories in assessing the identified factors. A total of 92 respondents in the Malaysian public sector participated in this study. Structural Equation Modelling with Partial Least Square is the main statistical technique used in this study. The study has revealed that human characteristics such as knowledge and innovativeness to EA and technological characteristics such as relative advantage and complexity of EA influence its implementation by the Malaysian public sector. Based on the findings, the theoretical and practical implications of the study as well as limitations and future works are also discussed.**

***Keywords*—Enterprise architecture; public sector; HOT-Fit**

I. Introduction

Enterprise architecture (EA) is a blueprint of the fundamental structures of an organization that describes the processes used for development of information technology (IT). The EA components and relations acts as an enabler for organization. EA aims to cover several issues and factors that determine organization to adopt EA. Many organizations have been implementing EA, but there are still many organizations have been unwilling to do so. This reluctance of organizations to implement EA contrasts with earlier forecasts by some of the proponents of EA that anticipated it to have been widely implemented by now. As far as EA implementation is concerned, literature and research into how organization continue with their strategy for enterprise architecture implementation is limited [1]. Moreover, study by [2] also pointed out that while extant research has studied EA, costs and benefits, potential applications, the influence factors of EA implementation in organization particularly in the public sector has not been empirically examined.

The phenomenon of EA implementation can be apprehended and measured by evaluating the organizational level of technology adoption. The phenomenon of technology adoption by organization can be defined as the adoption of an idea that is new to the organization adopting it. In this context, EA is conceptualized as a new idea of technology that related to the IT management to the organization to adopting and implementing it. Hence, many studies have developed models that explain or predict the adoption or implementation decision and extent of diffusion of technology within an organization. There are theories, models, theoretical frameworks and conceptual framework that have been developed to examine organizational technology adoption, characteristically dealing with decision to adopt, intention to adopt, intention to use, adoption, implementation and diffusion.

In addressing the aforesaid problems, this study explores and evaluates on critical success factors as well as evaluating the implementation of EA in the Malaysian public sector. Therefore, the contribution provided by the research presented in this paper is two-fold. First, we reveal and explain the influenced factors of EA implementation based on Human-Organization-Technology (HOT-Fit) Model by [3]. Second, we develop a model that will be used to measure the significance of the identified factors influencing the EA implementation success within the Malaysian Public Sector context.

II. Literature Review

A. *Enterprise Architecture in the Malaysian Public Sector*

In 2013, 1-Government Enterprise Architecture (1GovEA also known as MyGovEA starting in Jun 2018) has been announced by Malaysian Administrative Modernization and Management Unit (MAMPU). The purpose of MyGovEA is to strengthen and align ICT policy, standards and practices with the government vision and mission [4]. Moreover, the aim of MyGovEA is to assist the Malaysian public sector in aligning and unifying the business and IT strategy in order to meet the agency vision and mission towards better service delivery. The framework of MyGovEA as shown in Fig. 1, consist of three main components, namely, architecture domains, tools and repository, and methodology. Another supporting component is governance and principles.

MyGovEA architecture domains component is the core initiative for the Malaysian public sector EA. At the Business Architecture layer, capabilities and end-to-end business processes, functions, enterprise business outcomes, and their relationships to external entities required to execute business strategies is defined. For this component, the Malaysian public sector has initiated 1Malaysia Training Centre (1MTC also known as MyTC starting in Jun 2018) and 1Government Unified Communications (1GovUC also known as MyGovUC starting in Jun 2018). Meanwhile, for Information and Application Layers which deals with the structure and utility of information within the organization, and its alignment with its strategic as well as tactical and operational needs specifies the structure of individual systems based on defined technology. Thus, there are Public Sector Big Data Analytics (DSRA), Public Sector Data Dictionary (DDSA) and Public Key Infrastructure (PKI) initiatives that are initiated. Whereas, for technology layers, which defines the technology environment and infrastructure in which all IT systems operate, four initiatives were introduced which are Putrajaya Campus Network (PCN), 1Government Network (1Gov*Net also known as MyGov*Net starting in Jun 2018), Public Sector Data Centre (PDSA) and Malaysia Government Comprehensive Managed ICT Security Services (MyGSOC).

The four architecture domains largely represent the current state of practice in the discipline of MyGovEA. The successful of MyGovEA is not only captured by the four domains, but also the relationships between them. Having linkages between the four domains provides connection-of-sight to the relevant stakeholders of the EA.

B. *Critical Success Factors for Implementation of Enterprise Architecture*

A preliminary study was conducted to identify critical factors for successful implementation of EA. At this stage, a total of 239 papers that is matched with the keyword of "enterprise architecture" is selected from the various databases such as ScienceDirect, Wiley, JStor, EBSCO, Emerald, as well as InderScience. After screening process and filtering for selective criteria, only 16 papers selected for further meta-analysis process. The first paper used a meta-analysis of previous studies on critical success factors (CSF) for IT innovation adoption in government sector [5] as well as related to EA implementation in the government sector [6], [7] and [8]. Another paper used to identify CSFs was based on interview results where authors interviewed EA projects stakeholders that's include of an IT Manager and EA team members. After the meta-analysis process, the summary in the Table I reported critical success factors on the EA is classified based on HOT-fit themes:

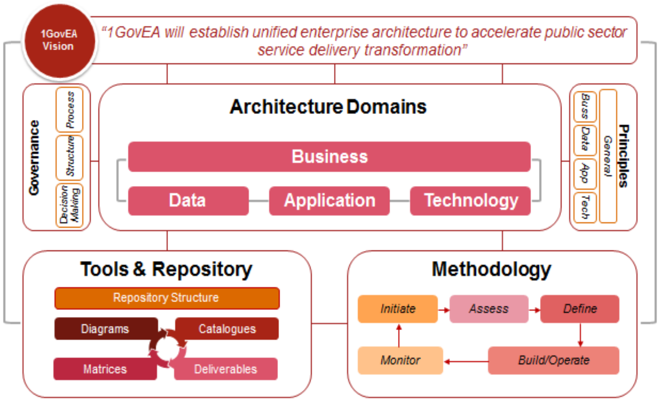


Fig. 1. MyGovEA for the Malaysian Public Sector (Source: www.mampu.gov.my).

TABLE I. Critical Success Factors of EA Implementation in the Public Sector based on HOT-Fit Framework

Context	Factors
Human	EA team competency Knowledge and Skills Teamwork Silo thinking Innovativeness
Organization	Change Management Human Resource Top management support Organizational Readiness Financial Governance Project Championship Managerial Capability Formalization
Technological	Supporting tools Complexity Fragmented systems Legacy systems Privacy and securityCost

III. Theoretical Foundation

The consideration of the IT systems implementation and adoption literature can be subdivided into two folds; individual and organizational level. In this paper the researchers conceptualized the EA implementation on the organizational level within the setting of the Malaysian public sector. Therefore, the HOT-fit model by [3] is utilized for the purpose of this study that fits concept between human, organizational and technological context. The following subsections describe the conceptualization of HOT-fit model as a foundation framework for EA implementation in the Malaysian public sector.

A. Human Factors

Along with the IT innovation literature, many distinctive variables are possible determinants of organizational adoption of innovation. One of the distinctive contexts is the characteristics of human who are involved with IT innovation adoption. With a focus on the human factors in IT innovation adoption study, the HOT-fit model that integrates human dimensions by [3] is suitable as a based framework to evaluate human context for EA implementation in the Malaysian Public sector.

Previously, [9] define IT adoption is determined by the characteristics of the individuals in the organization. Later, [10] in his study has included Chief Executive Officer (CEO) characteristics to his theoretical framework on IS adoption. He found that the characteristic of CEO is a distinct domain for SME to adopt IS.

Furthermore, [11] and [12] argued that IT professional participation to the IT activities in the organization requires skills and knowledge in technology, interpersonal as well as management to effectively integrate IT with business in the organization. Additionally, they also proposed that IT professionals must able to learn past, current and future IT development in organization to meet the evolving needs. Most IT professionals just hold their skills and knowledge on the previous IT technology without the motivation or innovativeness to explore and learn current trends of the IT. As a reality, IT professionals likewise needs to upgrade their knowledge and change their innovativeness with the changing of IT development, business needs and the scope of IT requirement in the organization. For that IT officer innovativeness and IT officer knowledge will be examined in this study to evaluate the affect to EA implementation by the Malaysian public sector.

1) IT officer innovativeness: As a norm in the Malaysian public sector that IT officer in the organization is considered as the owner, influencer or decision maker for organization to implement IT innovation. The involvement of IT officer in the organizations' IT project start with the proposal phase until the implementation phase. In the same systems such as an organization, other members relative to other member innovativeness to accept and adopt new ideas [13]. This idea is employed by [14] to apprehend CIO innovativeness in the direction of the adoption of recent information technology. Moreover, an innovativeness of IT officer is prepared to take the risks and typically choose new answers which have now not been tried previously. Latest literatures found that the significant IT officer innovativeness positively influence the adoption of IT innovations. Therefore, the following hypotheses were formulated:

H1: IT officer innovativeness will have positive effect to EA implementation in the Malaysian public sector.

2) IT officer knowledge: The knowledge in the context of this study states the technical knowledge about technologies by IT officer. The technical knowledge states the ability to perform all of the IT task required in the IT project such as technical capability, processes, tools and skills. The lack of technical knowledge, expertise and skills in by the IT officer contribute to the slow rate of IT innovation adoption. During the initiation phase, the organization gathers the necessary information about IT innovation for new IT project and the evaluation of the information deeply depends on the knowledge of IT officer. Thus, the capability of organization for implementing the new IT project such as EA, the necessary skilled and adequate technical knowledge of IT officer for adoption is required. This argument supported by the previous study on the adoption or human resource information systems (HRIS) in Singapore by [15] which found that the success and sustainable growth of HRIS because of the availability of technical knowledge by IT professional. Next, in the context of cloud computing implementation, [11] and [14] argued that IT officer must have a knowledge about the types of cloud computing such as deployment models and also able to choose the right models for organizations. Therefore, in the context of EA implementation, IT officer must have a knowledge about the concept, tools, as well as the process of EA implementation. Thus, the following hypothesis were formulated.

H2: IT officer knowledge will have positive effect to EA implementation in the Malaysian public sector.

B. Organizational Factor

The context of organizational factor in this study represents the organizational characteristic that is described as the organizational characteristics in the Malaysian public sector. For instance, the organizational characteristics that will determine the capability of the Malaysian public sector to implement and practice EA such as current IT readiness and the support of top management.

1) Organizations' IT readiness: IT readiness is conceptually defined by the three dimensions. It makes more sense to state that an organization has higher level of IT readiness because it possesses IT human resource, IT infrastructure and system integration than it does to state that the higher-level IT readiness leads to higher level IT human resource, IT infrastructure, and system integration [17]. The three dimensions form the overall IT readiness as a composite. IT human resource, infrastructure, and system integration reflects IT readiness from three different aspects respectively. The lack of one aspect may lead to incomplete evaluation of the overall IT readiness in a firm. These subdimensions of IT readiness are essentially distinct and independent from each other especially for EA implementation that consist of all these elements of IT readiness. For example, a comprehensive IT documentation policy and guidelines as well as IT infrastructure can provide standards for other IT components such as IT architecture (software, hardware, data exchange, communications, operating systems, etc.) and ensure the smooth transition of "as-is" to "to-be", which are required to implement EA. Thus, the following hypothesis were formulated.

H3: Organizational IT readiness has a significant relationship with the implementation of EA.

2) Top management support: In the big organization such as the public sector, the ultimate influencer for the new IT project or IT innovation implementation is the support from the top management [16], [18] and [19]. In this study, top management support refers to the extent to which top managements are involved in and promote the implementation of EA by the organization. Director or Head of Department are the individuals who are authorized and responsible to make a decision concerning strategic movements and resource allocation, which can also substantively impact implementation process and results of IT innovation within the organization. Top management can create a climate that favors EA implementation by motivating and thinking about process innovation with EA, overcoming organizational inertia or structural barriers, ensuring and continuing attention to EA initiatives and facilitating strategic renewal to align EA with business objectives. Therefore, the support from the top management is important for the new IT project like EA to be implemented. Subsequently, top management support creates supportive environment and impart an appropriate resource such as financial as well as human resource. Hence, the following hypothesis were formulated.

H4: Top management support has a significant relationship with the implementation of EA.

C. Technological Factor

The context of technological factors in this study represents the characteristics of EA tools that might determine the likelihood of EA implementation by the Malaysian public sector. This assertion is supported by the previous studies which is conceptualize the characteristics of technology such as cloud computing, e-business, mobile application, human resource systems, social media applications, enterprise resource planning, customer relationship management as well as enterprise architecture as an important factor for the implementation and adoption of such technologies [20], [21]. Moreover, the literature additionally recommended further research on the influences of the technological characteristics such as relative advantage and complexity for EA implementation in the public sector [4], [12].

1) Relative advantage: Relative advantage refers to the benefits of the technology can provide to the organization. The degree of the benefits evaluated in terms of productivity, profitability, time and effort, and cost. Recent study by [2] indicates that an organization will get a lot of benefits by implementing EA such as effective IT decision support, ensuring data integrity and security, facilitating data analytics as well as easing the setting and enforcing standards for “to-be” IT environment. In the context of the Malaysian public sector, the benefits of EA implementation (MyGovEA for instance) will create better opportunities for the agency to enhance their IT capabilities in their organization. This advantage expected that EA’s positively related to implementation of EA in the Malaysian public sector. Thus, the next hypothesis was formulated.

H5: Relative advantage has a significant relationship with the implementation of EA.

2) Complexity: Complexity of the innovation defined as the degree of difficulty to understand and use of that innovation [13]. In the context of EA, the complexity of tools such as ArchiMate and Microsoft Visio as well as other modelling developing tools. Moreover, the complexity of the EA management and process also an inhibitor to implementation of EA. These complexities create greater uncertainty for successful implementation of EA in the Malaysian public sector and increase the risk of the implementation process. Thus, the following hypothesis were formulated.

H6: Complexity has a significant relationship with the implementation of EA.

IV. Research Design

A. Research Samples

The sampling frame of this study was 730 various departments in the Malaysian public sector. Manager or Head of IT Department and senior IT officer from each department was selected as a respondent. The selection of the respondents based on the roles as IT officer because they are fully involved in the IT project as an owner, leader or members in their department.

B. Measures Operationalization

The dependent variable in this study is the implementation of EA in the Malaysian public sector and the independent variables are within the context human, organization, technology discussed above. All measurement items adopted from established and various existing literatures. The 5-point Likert scale ranging from 1 to 5 (strongly disagree to strongly agree) used to measure the statements in the questionnaire.

C. Data Analysis

Data analysis process involved in this study is descriptive analysis, measurement and structural model analysis. Statistical analysis package, SPSS and Structural Equation Modelling (SEM) Partial Least Square (PLS) used as a software for the data analysis process. The data screening process such as missing data, straight lining, normality test, as well as non-response bias was conducted before further data analysis process. Out of 200 questionnaires distributed, only 92 respondents selected for further analysis.

Next, descriptive analysis using SPSS 21.0 was conducted to get the demographic background of respondents followed with the measurement analysis of the questionnaires using SmartPLS 3.0. All measurements were analysed to measure the reliability and validity of the items of the questionnaire and variables. Finally, structural analysis was conducted to test the hypotheses and validate the research model.

V. Results

A. Demographic Backgrounds

The background of the respondents in this research is illustrated in Table II. General background of the respondents such as age, gender, education, working experience and job position were derived from the first part of questionnaire. The information from Table II shows that the majority of the respondents are mainly a senior of IT officer in the various departments in the Malaysian public sector. They also have working experience more than 5 years. Therefore, their response to the questionnaire in this study is meaningful and reliable to represent their department.

B. Measurement Model Analysis

For the measurement model analysis, measuring the reliability and validity of the item's questionnaires was conducted using SmartPLS. An average variance extracted (AVE) used to measure the reliability of the item questionnaires and represent as factor loading. The criteria or cut-off value for the items to be considered as reliable is 0.7 and above [22]. Next, Cronbach's Alpha (α) and composite reliability (CR) used as a measurement criterion to assess the reliability and internal consistence reliability of the variables. The required value to meet the criteria for reliability is 0.5 for Cronbach's Alpha and 0.7 for CR [22].

TABLE II. Descriptive Statistics of Demographic Backgrounds

Background	**Information**	**Frequency**	**Percentage %**
Gender	Male	51	55.4
	Female	41	44.6
Age	25 or less	12	13.0
	26 - 35	13	14.1
	36 - 45	30	32.6
	46 - 55	22	23.9
	56 - 60	15	16.3
Education Level	PhD	2	2.2
	Master Degree	11	12.0
	Bachelor Degree	53	57.6
	Diploma	12	13.0
	Others	14	15.2
Present Position	Chief Information Officer (CIO)	3	3.3
	Manager/Head of IT Agency	8	8.7
	IT Officer	68	73.9
	IT Staff	11	12.0
	Others	2	2.2
Working Experience in the Malaysian Public Sector	Less than 2 years	8	8.7
	Between 2-4 years	12	13.0
	Between 5-7 years	19	20.7
	Between 8-10 years	25	27.2
	More than 10 years	28	30.4
Agency Types	Federal	38	41.3
	Federal Statutory	15	16.3
	State	10	10.9
	State Statutory	16	17.4
	Local Authority	13	14.1

From Table III, the factor loading of the items above the value 0.5 which is ranging between 0.934 and 0.703 and satisfied the criteria for reliability test stated by Hair et al. 2010. The value of the AVE, α and CR also meet the criteria of the reliability test and higher than cut-off values for each criterion, thus the measurement of the items in this study is reliable.

Next, this study proceeded to test of the construct validity. The analysis used to validate the data and variables in this study is discriminant validity. In this test, the evaluation is done by comparing between correlation value and square root of AVE for each variable. The result in the Table IV shows that the value for inter-correlation is lower than the value of square roots of AVEs. Thus, the validity for measurement instruments in this study is validated. Fig. 2 illustrates the measurement model of this study.

C. Structural Model Analysis

The next analysis performed for this study was the evaluation of the hypotheses. In this analysis, all hypotheses in this study represent by the arrow relationship from independent variables to the dependent variable in the SmartPLS model. In this study, the method first order and second order construct was applied as suggested by [23]. There are two level of independent variables conceptualized as a second order and first order construct. For the first order, all 6 independent variables form a first order construct. For the second order construct, human, organizational and technological form as combination for their first order variables. For instance, second order construct of human is representing by combining all items from knowledge and innovativeness. Similar to organizational and technological construct which formed by the combination of their first order constructs.

TABLE III. Quality of the Measurement Model

Latent Variable	Items	Loading	AVE	CR	CA
Knowledge	KNW1	0.8265	0.7038	0.8769	0.7893
	KNW2	0.8667			
	KNW3	0.8229			
Innovativeness	INNOV1	0.9276	0.8155	0.9298	0.886
	INNOV2	0.9340			
	INNOV3	0.8448			
Top Management Support	TMS1	0.7720	0.6605	0.8529	0.7386
	TMS2	0.9016			
	TMS3	0.7567			
Organizational Readiness	ORS1	0.8885	0.8081	0.9266	0.8805
	ORS2	0.9394			
	ORS3	0.8673			
Relative advantage	RA1	0.8247	0.7078	0.879	0.7939
	RA2	0.8312			
	RA3	0.8674			
Complexity	CPLX1	0.8128	0.6973	0.8733	0.7813
	CPLX2	0.8891			
	CPLX3	0.8004			
EA Implementation	IMPL1	0.8009	0.5712	0.8417	0.7634
	IMPL2	0.7030			
	IMPL3	0.7586			
	IMPL4	0.7574			

Note: AVE = Average variance extracted; CR = Composite Reliability; CA = Cronbach Alpha

Next, model analysis was performed to examine the significant of the relationship between independent variables and dependent variable. This test also known as hypotheses or the value of the t-value. Bootstrapping process was initiated to generate the t-value of the model analysis. The total number of 5000 samples and two-tailed criteria used (t > 1.67) in this study as suggested by [24].

TABLE IV. DISCRIMINANT VALIDITY OF LATENT VARIABLES

		1	2	3	4	5	6	7
1	Knowledge	**0.8389**						
2	Innovativeness	0.6624	**0.9031**					
3	Top Management Support	0.2085	0.2105	**0.8127**				
4	Organizational Readiness	0.3294	0.3631	0.5469	**0.8989**			
5	Relative advantage	-0.017	0.1011	0.3320	0.3317	**0.8413**		
6	Complexity	0.4254	0.4169	0.2268	0.2250	0.1314	**0.8350**	
7	EA Implementation	0.0861	0.3362	0.1808	0.0861	0.3784	0.322	**0.7558**

TABLE V. RESULTS OF HYPOTHESIS TESTING

	Hypotheses	Path Coefficient	*T* value	*P* value	Decision
H1	INNOV -> EA IMPL	0.1777	2.392	0.02	SUPPORTED
H2	KNOW -> EA IMPL	0.1515	2.292	0.02	SUPPORTED
H3	ORS READINESS -> EA IMPL	-0.0931	1.014	0.31	NOT SUPPORTED
H4	TMS -> EA IMPL	-0.0719	1.048	0.30	NOT SUPPORTED
H5	RA -> EA IMPL	0.3070	2.331	0.02	SUPPORTED
H6	COMPLEXITY -> EA IMPL	0.2703	2.723	0.01	SUPPORTED

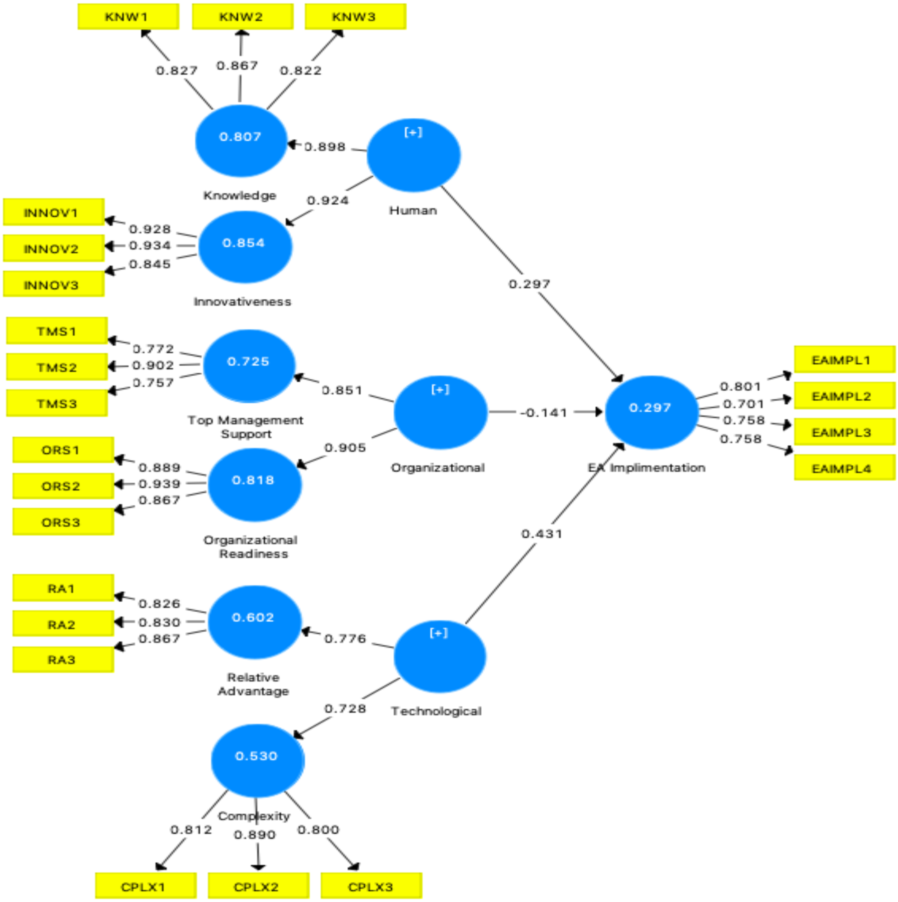


Fig. 2. Measurement Model.

The result of the structural model analysis in this study is shown in Table V. The result shows that the t-value of relationship between knowledge, innovativeness, relative advantage and complexity is greater than 1.67, thus H1, H2, H5 and H6 significantly affect to EA implementation in the Malaysian public sector. The results also revealed that both total effects for organizational readiness and top management support not significant to EA implementation by the Malaysian public sector, thus both hypotheses H3 and H4 is rejected.

The last step of structural model analysis is the evaluation percentage variance or R^2 value of dependent variable. The value of R^2 represents the percentage of variation explained by the independent variables to predict the accuracy of the model. The higher value of R^2, the higher predictive accuracy of the independent variables toward dependent variable. According to [25], the model having R^2 as 0.67 and above considered substantially accurate. While the model that having the value of R^2 below than 0.19 are considered as weak accuracy. The accuracy of moderate if the value of R^2 is between 0.19 and 0.67. In this study, the result of R^2 value for EA implementation is 0.30, thus can be consider moderate accuracy predicted by human, organizational and technological variables.

VI. Discussion

The objective of this study is to identify and evaluate the factors that influence the implementation of EA by the Malaysian public sector. After conducting the process of preliminary research, HOT-fit model was selected as a foundation theory to develop the model for this research to determine the factors in the context of human, organizational and technological. Along with the research process and analysis conducted, the results show that both human and technological factors positively influence the implementation of EA by the Malaysian public sector. These results in line with the previous research in their study in other IT innovation adoption and implementation [26], [27] and [28]. These results suggest that, the factors of IT officer knowledge and innovativeness toward EA positively influence the department in the Malaysian public sector to implement EA.

Next, the results of this study also show that technological factors such as relative advantage and complexity influence the implementation of EA by the Malaysian public sector. The results of this study confirm the findings by the previous study within the context of other IT innovation adoption and implementation [29]. The benefits of EA implementation were informed by the IT officer as well as the complexity of process and EA tools. This factor also inter-correlated with the knowledge of EA by the IT officer. From the knowledge gain, IT officer proves that EA is beneficial to be implemented by their department.

However, the results of this study found that organizational factor had been insignificant to the EA implementation by the Malaysian public sector. The role of the organization in influencing the team on the EA implementation seems to be insignificant. The support of the top management and IT readiness by the department seems to be not the factors to the EA implementation. This can be argued that, the structure of administrative level of the department in the Malaysian public sector might factors. The respondents might be confused with the support of the top management in this study either their decision maker within the department or the decision maker among the ministry level. Departments in the public sector also serve with different level of decision maker. Some department receives direction from the ministry level and some of the departments receive the direction from their Head of Department. In terms of IT readiness, in can be argued that the implementation of EA does not require the support of IT infrastructure and resources. The process of EA implementation is involved with management process and little features on EA tools. Thus, the requirement of complete IT infrastructure and human resource seems to be not required.

VII. Conclusions

This paper proposes to develop an extended EA implementation model for the Malaysian public sector that is composed of HOT-fit such as human, organizational and technological factors as independent variables and EA implementation as dependent variable. All four independent variables, IT officer knowledge and innovativeness on EA as well as relative advantage and complexity of EA significantly influence the EA implementation. However, two independent variables, top management support and IT organizational readiness not influence the EA implementation. Thus, the results of this study contribute to the body of knowledge in the context of HOT-fit model and EA.

After conducting the analysis based on quantitative method, this study provides a view for the important factors that might play crucial roles in the EA implementation process in the Malaysian public sector. However, this study needs to be further validated using qualitative method either by the case study or details interview with the main player of EA in the Malaysian public sector.

Acknowledgements

We gratefully acknowledge that this research paper was funded by the National University of Malaysia Research Grant GGPM-2018-012. We would also like to acknowledge the editors and anonymous reviewers for their valuable guidance.

References

[1] F. Nikpay, R. Ahmad, B. D. Rouhani, and S. Shamshirband, "A systematic review on post-implementation evaluation models of enterprise architecture artefacts," Inf. Syst. Front., pp. 1–20, 2016.

[2] N. A. A. Bakar, N. Kama, and S. Harihodin, "Enterprise architecture development and implementation in public sector: The Malaysian perspective," J. Theor. Appl. Inf. Technol., vol. 88, no. 1, pp. 176–188, 2016.

[3] M. M. Yusof, J. Kuljis, A. Papazafeiropoulou, and L. K. Stergioulas, "An evaluation framework for Health Information Systems: human, organization and technology-fit factors (HOT-fit)," Int. J. Med. Inform., vol. 77, no. 6, pp. 386–398, 2008.

[4] N. A. A. Bakar, S. Harihodin, and N. Kama, "Assessment of Enterprise Architecture Implementation Capability and Priority in Public Sector Agency," Procedia Comput. Sci., vol. 100, pp. 198–206, 2016.

[5] M. M. Kamal, "IT innovation adoption in the government sector: Identifying the critical success factors," J. Enterp. Inf. Manag., 2006.

[6] M. M. Kamal and M. Themistocleous, "Investigating EAI adoption in the local government authorities: A case of mapping the influential factors on the adoption lifecycle phases," Transform. Gov. People, Process Policy, vol. 3, no. 2, pp. 190–212, 2009.

[7] Y. Abd.Rahim, M. Mohamad, N. Safie, and Z. And Rahim@Ab Rasim, “Faktor Kejayaan Kritikal , Cabaran Serta Kebaikan Pelaksanaan Seni Bina Perusahaan (EA) Dalam Agensi Awam Malaysia,” MALTESAS Multi-Disciplinary Res. J., vol. 3, no. 2, pp. 62–71, 2018.

[8] N. A. A. Bakar, M. N. Kama, and H. Selamat, “Service-Oriented Enterprise Architecture (SoEA) Adoption and Maturity Measurement Model: A Systematic Literature Review,” Int. J. Comput. Electr. Autom. Control Inf. Eng., vol. 7, no. 12, pp. 334–345, 2013.

[9] J. Y. Thong and C. S. Yap, “CEO characteristics, organizational characteristics and information technology adoption in small businesses,” Omega, vol. 23, no. 4, pp. 429–442, Aug. 1995.

[10] J. Y. Thong, “An integrated model of information systems adoption in small businesses,” J. Manag. Inf. Syst., vol. 15, no. 4, p. 187, 1999.

[11] H. Sallehudin, R. C. Razak, and M. Ismail, “Factors Influencing Cloud Computing Adoption in the Public Sector: An Empirical Analysis,” J. Entrep. Bus., vol. 3, no. 1, pp. 30–45, 2015.

[12] H. Sallehudin, R. C. Razak, M. Ismail, A. F. M. Fadzil, and R. Baker, “Cloud Computing Implementation in the Public Sector: Factors and Impact,” Asia-Pacific J. Inf. Technol. Multimed., vol. 7, no. 2–2, pp. 27–42, 2019.

[13] E. M. Rogers, Diffusion of innovations. 5th edition, New York: Free Press., 2003.

[14] R. Agarwal and J. Prasad, “A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Technology,” Inf. Syst. Res., vol. 9, no. 2, pp. 204–215, Jun. 1998.

[15] T. S. H. Teo, G. S. Lim, and S. A. Fedric, “The adoption and diffusion of human resources information systems in Singapore,” Asia Pacific J. Hum. Resour., vol. 45, no. 1, pp. 44–62, Apr. 2007.

[16] H. Sallehudin, R. C. Razak, and M. Ismail, “Determinants and Impact of Cloud Computing Implementation in the Public Sector,” J. Adv. Inf. Technol., vol. 7, no. May, pp. 245–251, 2016.

[17] C. Y. Lin and Y. H. Ho, “Determinants of Green Practice Adoption for Logistics Companies in China,” J. Bus. Ethics, vol. 98, no. 1, pp. 67–83, Jun. 2010.

[18] J. Y. Xin, T. Ramayah, P. Soto-Acosta, S. Popa, and T. Ai Ping, “Analyzing the Use of Web 2.0 for Brand Awareness and Competitive Advantage: An Empirical Study in the Malaysian Hospitability Industry,” Inf. Syst. Manag., vol. 31, no. 2, pp. 96–103, Apr. 2014.

[19] S. M. Motahar, M. Mukhtar, N. S. Mohd Satar, M. Y. Maarif, and S. Mostafavi, “Revisiting the Diversification on the Implementation of Open Source ERP Teaching Models,” Jour Adv Res. Dyn. Control Syst., vol. 10, no. 09, pp. 2379–2385, 2018.

[20] A. Mukred, D. Singh, and N. Safie, “Investigating the impact of information culture on the adoption of information system in public health sector of developing countries,” Int. J. Bus. Inf. Syst., vol. 24, no. 3, p. 261, 2017.

[21] Y. A. Rahim, M. Mohamad, N. Safie, A. Rahim, and A. Rasim, “Critical Success Factors, Challenges and Benefits of Enterprise Architecture (EA) in the Malaysian Public Agency,” MALTESAS Multi-Disciplinary Res. J., vol. 3, no. 2, pp. 62–71, 2018.

[22] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, Multivariate Data Analysis (6th ed.). 2006.

[23] P. B. Lowry and J. Gaskin, “Partial least squares (PLS) structural equation modeling (SEM) for building and testing behavioral causal theory: When to choose it and how to use it,” IEEE Trans. Prof. Commun., vol. 57, no. 2, pp. 123–146, 2014.

[24] J. F. . Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). Thousand Oaks, CA: Sage Publications, Inc., 2014.

[25] W. W. Chin, “Issues and opinion on structural equation modeling,” MIS Q., vol. 22, no. 1, pp. vii–xvi, 1998.

[26] L. Ahmadian, S. Salehi Nejad, and R. Khajouei, “Evaluation methods used on health information systems (HISs) in Iran and the effects of HISs on Iranian healthcare: A systematic review,” Int. J. Med. Inform., vol. 84, no. 6, pp. 444–453, 2015.

[27] A. Meri et al., “Modelling the utilization of cloud health information systems in the Iraqi public healthcare sector,” Telemat. Informatics, vol. 36, no. April 2018, pp. 132–146, 2019.

[28] M. Mohammadi and M. Mukhtar, “A service-oriented methodology for supporting business process architecture layers in supply chain management,” Int. J. Inf. Syst. Supply Chain Manag., vol. 10, no. 4, pp. 18–43, 2017.

[29] M. Ghobakhloo and T. S. Hong, “IT investments and business performance improvement: the mediating role of lean manufacturing implementation,” Int. J. Prod. Res., no. June, pp. 1–18, Apr. 2014.

Robot Arm Analysis based on Master Device Pneumatic Actuators

Mohd Aliff[1], Nor Samsiah Sani[2]
Instrumentation and Control Engineering, Malaysian Institute of Industrial Technology, Universiti Kuala Lumpur, Johor[1]
Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia (UKM)[2]

***Abstract*—Advances in technology have expanded the use of soft actuators in various fields especially in robotics, rehabilitation and medical field. Soft actuator development provides many advantages, primarily being simple structures, high power to weight ratio, good compliance, high water resistance and low production cost. However, most soft actuators suffer the problem of being oversized which could potentially hurt users as it is often made of hard materials such as steels and hard rigid plastics. Current drawbacks of soft actuator implementation in robotic arms are on its excessive weights, causing these robots to be difficult to set up by patients themselves which in turn makes it less applicable for home rehabilitation training program. Hence, there is a need to design a soft actuator which is safe and more flexible, especially for applications in areas of patients in rehabilitation area or in-house rehabilitation program. In this paper, we propose the design of robot arm using master device pneumatic actuator and analyse the implementation for the above purpose. The system comprises primarily of the master and slave arms, two accelerometers and two potentiometers providing references for attitude control, six quasi-servo valves, and SH-7125 microcontroller. Our proposed design exhibits functions of the actuator that has been generated from elastic deformation of extension and contraction of the cylinder structure when high pneumatic pressure is supplied to the chamber. The control performance of the device is investigated using simulation method, whereby the rational model of the robot arm and the quasi servo valve with the embedded controller is implemented and analysed. It is found that the analysed results of the model approved well with the desired values.**

Keywords*—Trajectory control; master-slave control; robot arm; pneumatic cylinder

I. Introduction

A major social problem in developed countries is the alarming rise of elderly population against a continuingly declining birth rate. Such population differences have become more apparent every year, raising the concerns of various professional community from aspects of engineering, healthcare and social sciences. Today, many studies have shown a wide adoption of robot technology used in manufacturing [1], agriculture [2], communication [3], and education [4, 5]. In the line of 4IR, there is demand for one system that can control, communicate and integrate different robots regardless of their types and specifications. As term machine learning has heated up interest in robotics has not altered much. Only a portion of recent development in robotics can be credited to development any use machine leaning. Only a portion of recent development in robotics can be credited to development any use machine leaning [6]-[10]. Recent robotic development project has embedded machine learning algorithms to increase the intelligence in robots. This will increase the productivity while reducing the cost and electronic waste in a long run.

Specifically in the healthcare industry, robots are used in range of ways from supporting nursing care [11], assist the daily activities for elderly residents [12] as well as the disabled [13], to performing complex surgical procedures in hospital’s surgery situation [14]. The ability to perform work accurately and continually for extensive times is of significant merit in this scope. However, one major drawback of such robotic arms is that they are often heavy and need to be mounted on walls and pillars. This causes the room to be restricted in space and movement which thoroughly affects patients’ negatively. Current robot arm products are also complicated to be installed by patients themselves and are not applicable for home rehabilitation programs [15].

In medical and rehabilitation areas, several key features for actuators are expected; light weight, portability and adaptability to expand to needs, as well as usability in various environments [16, 17]. For these reasons, choosing the right actuator is important to satisfy the above criteria. A flexible and lightweight actuator has been developed in this study and applied for robotic arm to be used as rehabilitation device. In specific, the device is a novel type of flexible pneumatic actuator with two important attributes: (i) perform well even if the actuators are impaired by the extrinsic forces [18, 19] and (ii) flexibility of the robot arm based on simple structure, which is proposed and tested using the flexible pneumatic cylinders [20, 21]. This robot arm possesses three special degrees-of-freedom characteristics which are able to bend, extend and contract. With the aim to be harnessed in physiotherapy, the master and slave device is incorporated into the robot arm, to assume the situations where the physical therapist performs movement treatment to a patient. During rehabilitation, the therapist control the robot movement by holding the master arm and move it repeatedly according to conditions and needs of the patient. Simultaneously, the user’s arm placed on slave arm will actually mimic the master arm movement as presented in Fig. 1. Using a flexible pneumatic cylinder, the analysis of robot arm for the human wrist rehabilitation is introduced in this paper. The system consists of a set of master and slave arms, two accelerometers and two potentiometers providing references for attitude control, six quasi-servo valves, and SH-7125 microcontoller. Next, the analytical model of the flexible pneumatic cylinder and the quasi servo valve with the embedded controller was designed. A comparison of the theoretical and experimental results was then performed to verify the rationality of the designed model.

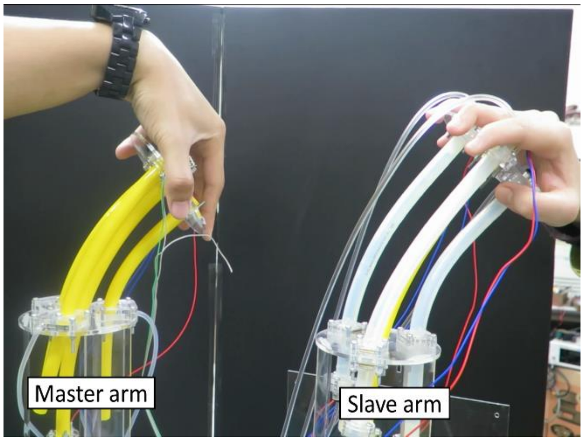


Fig. 1. Master and Slave Device of Robot Arm.

II. Material and Method

A. Structure of Robot Arm

Construction of flexible soft tube is shown in Fig. 2. Main components of the soft tube include brass rollers, flexible tubes, slide stage as well as 9 mm and 3 mm of steel balls. Soft polyurethane tube from SMC Co Ltd (TUS 1208) has been used as a cylinder that allows air to pass through it. When air is supplied into the tube, two pairs of brass rollers from top and bottom sides are pinched on the 9 mm steel ball (referred as the cylinder head), allowing free movement of the steel ball. On the other hand, a 3mm diameter steel ball was introduced in the middle of the slide stage and flexible tube to withstand the cylinder and brass rollers, thus enabling a natural tube movement. In addition, steel ball (9 mm) located amidst the slide stage was pushed and moved appropriately when pressure was given from one end of the flexible tube. In parallel, brass rollers also pushed the steel ball, moving the slide stage although it impaired the tube. Therefore, this ideal combination between center distance (*D*) and distance (*W*) between two couples of rollers can be a variable to find lowest handling pressure of the pneumatic cylinder. As a result, the best value for *D* was indicated at 14.4 mm and *W* at 10 mm after considering that this value combination caused the steel ball (9 mm) to not being able to escape from the slide stage and slender frictional force of the slide stage [18].

Table I highlights characteristics of flexible cylinder used in this study. Using the developed flexible pneumatic cylinders, the flexible robot arm is constructed as shown in Fig. 3 [20, 22]. The robot arm size specification is as follows: an overall diameter of Ø100 mm x 250 mm with a total mass of 380 g. It is partitioned as upper and lower round stages, fixed with a central tube, three flexible pneumatic cylinders, an accelerometer, three slide stages and a potentiometer. The preliminary distance between two round stages is about 100 mm after being measured using the potentiometer from Copal Electronics that was connected to the middle of the tube. The mid angle of two nearby slide stages is equal to 120 degrees on the stage following each flexible pneumatic cylinder's arrangement after its end was set to the upper stage. One of flexible pneumatic cylinder was driven by two quasi-servo valves [23, 24] which consist of four on/off valves (Koganei Co. Ltd., G010HE-1). Six control valves were needed to control the three flexible cylinders. In conditions that exerted 500 kPa amount pressure; the maximum bending angle of the robot arm was at 45 degrees with the overall produced force of three cylinders to be that of 45N. The operating principle for pneumatic cylinder is shown in Fig. 2. To bend the pneumatic cylinder to the right side, pressure must be directed to both ends of the right cylinder whereas the other two cylinders are pressurized from the top [17]. On the other hand, one end of the cylinder was pressurized in order to move the robot arm upward or downward.

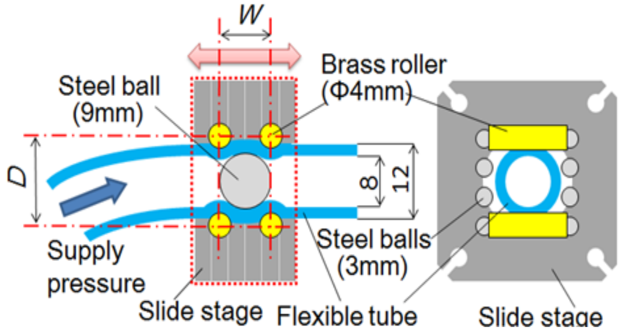


Fig. 2. Construction of the Flexible Pneumatic Cylinder.

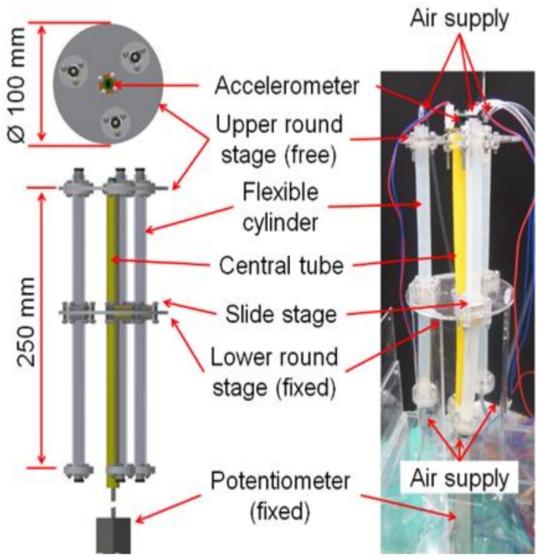


Fig. 3. Flexible Pneumatic Robot Arm.

TABLE. I. Characteristics of Flexible Cylinder

Weight	< 0.1kg
Pushing and pulling speed	> 1m/s
Produced force	16N(input:500kPa)
Lowest handling pressure	120kPa
Maximum operational pressure	600kPa
Minimum curvature radius	about 30mm
Movement	Push-pull actions
Effective temperature	From -20 to +60 deg.C

B. Master Slave Control

The geometrical relationship of robot arm is shown in Fig. 4. Once arms are curved, the form of flexible cylinder from upper stage to lower stage is presumed to always be in round arc. From the center of the robot arm, angle α is defined as the bending angle from X axis while angle β denotes from Z axis to the normal vector at the centre. Following the counter clockwise starting from X axis, the cylinder length (displacement) of flexible pneumatic for the cylinder 1 located exactly on the X axis is labelled as $L1$, followed by $L2$ and $L3$ for cylinder 2 and cylinder 3 respectively.

Fig. 4 shows the geometrical relationship of the robot arm, producing the following obtained Equation (1):

$$R = \frac{L}{\beta} \tag{1}$$

Furthermore, from Fig. 4, the equations for $L1$, $L2$ and $L3$ (length of the cylinder between the upper surface of the stage and lower stage) are acquired and shown in Equations (2), (3) and (4), respectively.

$$L_{1i} = (R_i - r \cdot \cos\alpha_i) \cdot \beta_i \tag{2}$$

$$L_{2i} = \left\{R_i - r \cdot \cos\left(\frac{2\pi}{3} - \alpha_i\right)\right\} \cdot \beta_i \tag{3}$$

$$L_{3i} = \left\{R_i - r \cdot \cos\left(\frac{4\pi}{3} - \alpha_i\right)\right\} \cdot \beta_i \tag{4}$$

From the above equations, r is set at 33 mm measured as the radius of the round stage to the centre of slide stage in the cylinder. Subscripts i = m,s indicates the master arm (desired value) and the slave arm (present value), respectively. Subscript number (1, 2 and 3) indicates the location number of the cylinder. By referring to Equations (1) to (4) and based on the displacement of the master and slave cylinder, the control system can be performed. The accelerometer (Kionix KXR94-2050) has been used to measure the bending direction angle α and the bending angle β of robot arm. This consists of a mass, a spring and a capacitance type displacement sensor. Fig. 5 shows an analytical model to calculate the value of α, β, and the coordinate Xc, Yc, Zc of robot arm end.

$$\alpha = \cos^{-1}\frac{V_x}{\sqrt{V_x^2+V_y^2}} \tag{5}$$

$$\begin{aligned} \beta &= \cos^{-1}\left(\frac{V_z}{V_{z\,max}}\right) \\ X_c &= R\cdot(1-\cos\beta)\cos\alpha \\ Y_c &= R\cdot(1-\cos\beta)\sin\alpha \\ Z_c &= R\cdot\sin\beta \end{aligned} \tag{6}$$

Vzmax, *Vx*, *Vy*, and *Vz* values signify differences of *Vz* between the values in horizontal and vertical planes and the output voltages from the accelerometer in *X*, *Y*, *Z* axis, respectively. The length of cylinders for every bending state of the robot arm can be calculated using (Eq. 1) to (Eq. 6). From our experiments, the calculated angle shows good agreement to the measured value in which the error between the calculated angles and the measured error angles is found to be less than 1 degree [22, 24].

C. Control Procedure

Fig. 6 shows the view of the master-slave control system and its schematic diagram which consists of a slave arm and a master arm.

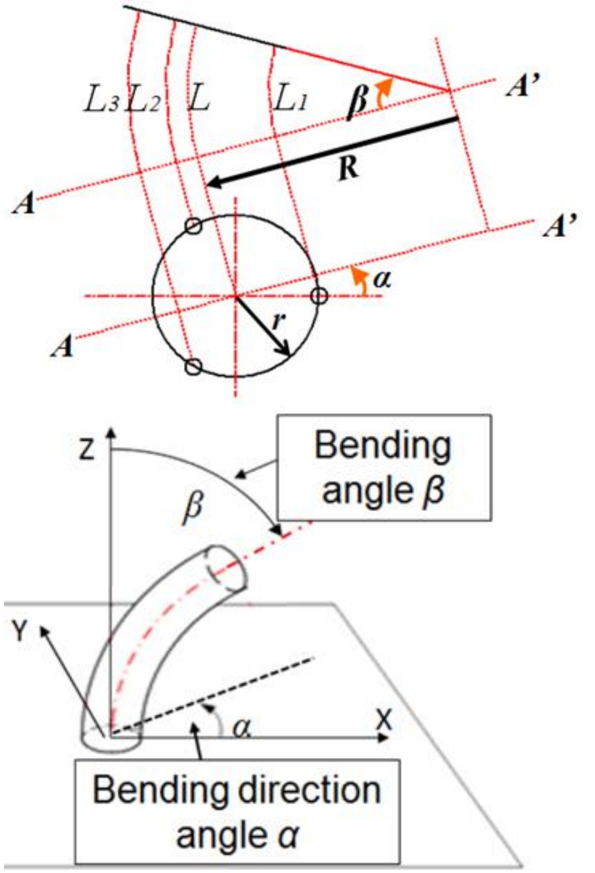


Fig. 4. The Geometrical Relationship of the Robot Arm.

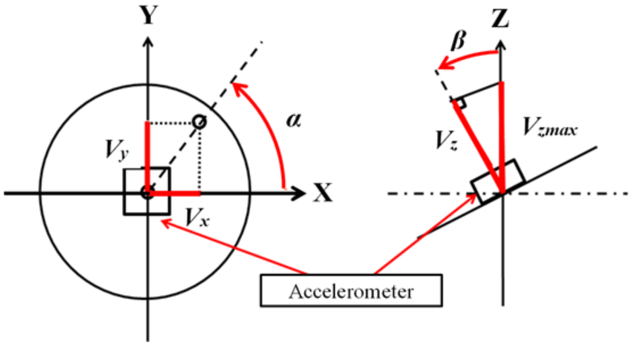


Fig. 5. The Correlation of Accelerometer Sensor with Bending Angle.

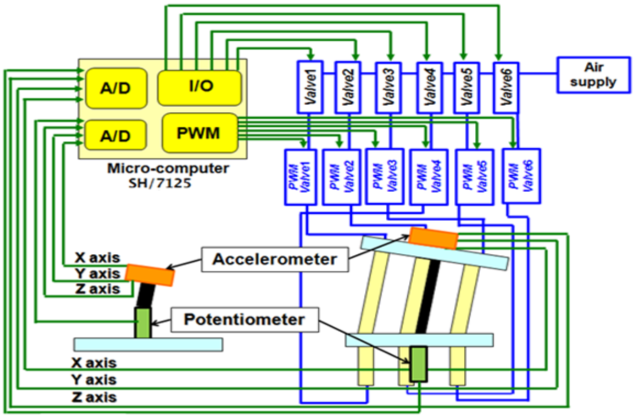


Fig. 6. Schematic Diagram for Control System.

In order to obtain the reference attitude value of the master arm, the accelerometer is set atop the upper stage. The changes of gravity for *X*, *Y*, and *Z* axes, the bending angle of upper stage of master and slave arm are then detected. Output voltages of the accelerometer from master and slave arm are sent to the microcomputer as shown in Fig. 6. Sampling period for master slave control system is set as 2.3 ms. The proposed control procedure for master-slave is described below:

1) Initially, the attitude of the master arm; the bending direction angle αm, the bending angle *βm* and the distance L_m are detected using the accelerometer and the potentiometer installed on it.

2) Distances of the master arm L_{1m}, L_{2m} and L_{3m} are calculated by the microcomputer using the analytical model from (Eq. 1) to (Eq. 4).

3) Attitude of the slave arm; *αs*, *βs* and *Ls* are next detected by the accelerometer and the potentiometer of the slave arm. Distances L_{1s}, L_{2s} and L_{3s} are calculated by using the analytical model.

4) Errors between L_{jm} for the master arm and L_{js} (j=1, 2, 3) for the slave arm are calculated using the microcomputer.

5) Finally, using both control system which are PWM control valve (quasi-servo valve) and PID control scheme depend on the calculated errors, the position of the cylinder can be controlled.

Another control system introduced in this paper is the trajectory control system. Such system exhibits only a slave arm. The procedure for trajectory control system using the analytical model is further described below:

1) X_d(t), Y_d(t), Z_d(t) is generated.

2) The present bending direction angle α, the bending angle β and the length L is determined using equations (5) and (6). The current arm end's coordinates X_c, Y_c, Z_c are obtained using these calculated values. Concurrently, the present cylinder lengths L1, L2 and L3 are also calculated using Equations (1) to (4).

3) X_{c1}, Y_{c1}, Z_{c1} of next arm end's coordinates are deliberated from the target trajectory. The distinctions δX, δY, δZ of current arm end's coordinates are obtained using the following equations:

$$\delta X = X_{c1} - X_c$$
$$\delta Y = Y_{c1} - Y_c \tag{7}$$
$$\delta Z = Z_{c1} - Z_c$$

4) Next, distinctions (δ_{Xc}, δ_{Yc}, δ_{Zc}) of arm end's coordinates are calculated based on the divergence between the preferred position of arm end and its current position. α_1, β_1 and L_1 are obtained using the following equations for which the preferred positions are calculated.

$$\alpha_1 = \alpha + \delta\alpha$$
$$\beta_1 = \beta + \delta\beta \tag{8}$$
$$L_1 = L + \delta L$$

5) The preferred cylinder lengths L_{11}, L_{21} and L_{31} are determined based on calculations of α_1, β_1 and L_1 from Equations (8) and Equation (1) to (4).

6) Finally, using PID control system and recapping the procedures from 2) to 5), variation of each cylinder length from the desired length is ascertained and position control of each cylinder is executed as below:

$$e_1 = L_{11} - L_1$$
$$e_2 = L_{22} - L_2 \tag{9}$$
$$e_3 = L_{33} - L_3$$

III. Result and Discussion

To validate the efficiency of the proposed model and identify system parameters, we evaluate the calculated result with the analysed result of the analytical model for the whole robot arm. As stated earlier, the following PID control scheme is embedded into the microcomputer as control modes:

$$u_i = \left| K_P e_i + K_I \int e_i dt + K_D \frac{de_i}{dt} \right| + 22.5 \qquad (i=1\sim3) \tag{10}$$

$$e_i = L_{im} - L_{is} \qquad (i=1\sim3) \tag{11}$$

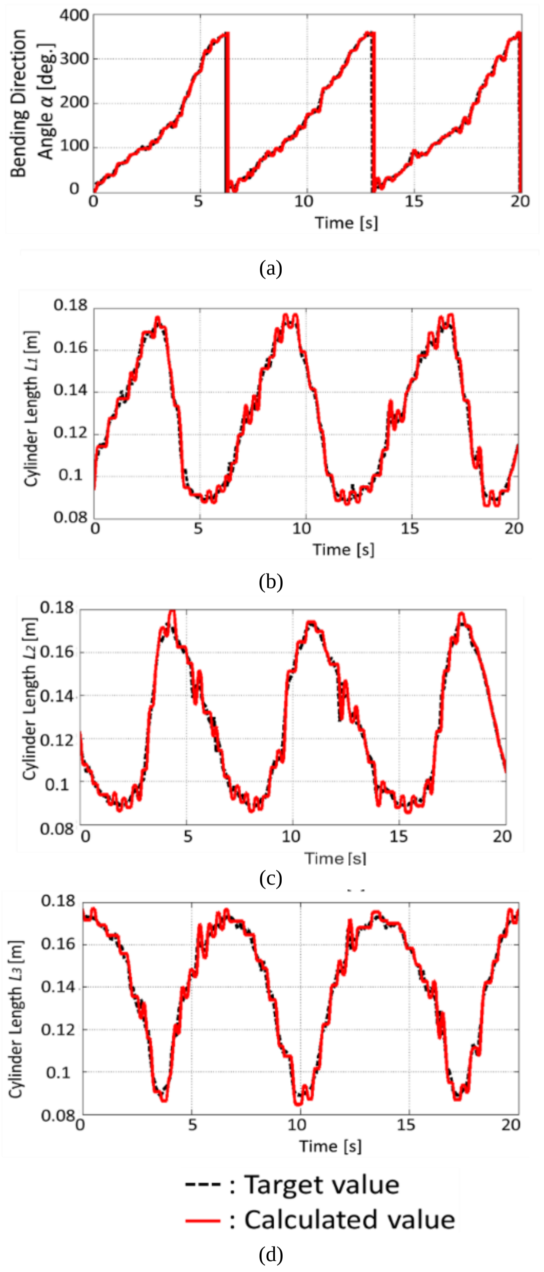
Where e_i[m] and u_i[%] indicate error ratio of the cylinder displacement and the duty ratio for the PWM valve in the quasi-servo valve, respectively. Table II shows the PID control parameters used for the simulation. By utilizing these selected values, the control error is able to become smaller while at the same time allowing the movement of the robot arm to be smoother.

TABLE. II. PID Control Parameter

	KP [%/mm]	KI [%/(mm·s)]	KD [%·s/mm]
Master-slave control	0.79	1.8	0.01

D/A converter has been developed to convert the variables output of microcomputer from digital into analog signal. The variables output voltages signify bending direction angle α and β and the displacement of the cylinders, in which these are logged using a GRAPHTEC, GL200 recorder.

Fig. 7 and 8 show the calculated results of the master-slave control and trajectory control, respectively. The target values which are calculated using the proposed equations are indicated as dashed black lines based on the accelerometer's output voltages of the master arm while the calculated results are indicated as the continuous red lines. From the evaluation between the target values and the calculated values, it can be confirmed that both of the results do match.



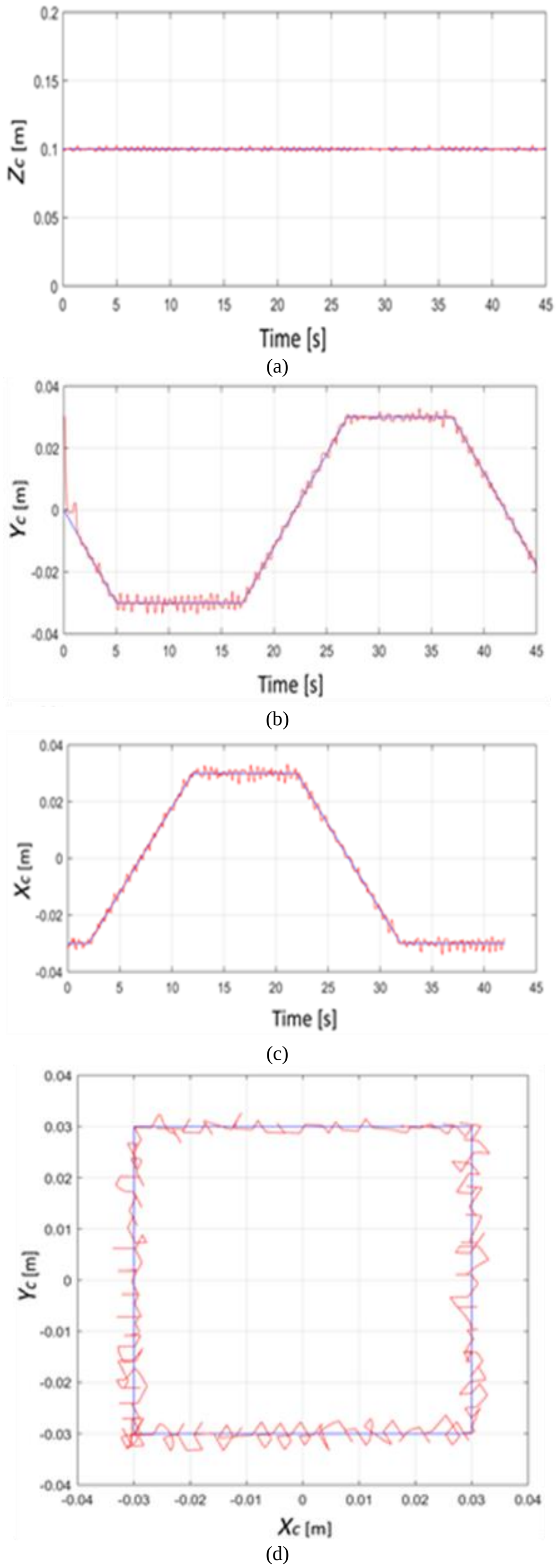
(a)

(b)

(c)

(d)

Fig. 7. Master-Slave Control Analysis; (a) Bending Direction Angle α Versus Time , (b) Cylinder Length, *L1* Versus Time, (c) Cylinder Length, *L2* Versus Time , (d) Cylinder Length, *L3* Versus Time.



(a)

(b)

(c)

(d)

Fig. 8. Trajectory Control Analysis (a) Z_c Versus Time,(b) Y_c Versus Time, (c) X_c Versus Time,(d) Y_c Versus X.

IV. Conclusions

This study's purpose is to develop a simple, light-weight and compact rehabilitation device for human wrist that can be set up and installed by patients with less complications. It should also be applicable for home rehabilitation training program. The system is shown to consist of inexpensive quasi-servo valves, a microcomputer, pneumatic robot arm and position accelerometers. Therefore, the development of flexible pneumatic robot arm for the master-slave control and trajectory control are proposed and investigated.

The simulation of the controls was performed based on our proposed analytical model. The values for PID control gain *KP*, *KI* and *KD* were investigated to achieve smaller control error, while at the same time enabling the robot arm to produce smooth movement. From our analysis and performance simulations, the calculated results agree with the desired values. From the comparison between the target values and the calculated values, we confirm that both of the results matched.

Acknowledgment

The authors would like to thank the Universiti Kebangsaan Malaysia and Ministry of Education Malaysia (MOE) under the Research University Grants (project code: GGPM-2017-039 and FRGS/1/2018/ICT02/UKM/02/6) for funding and supporting this research.

References

[1] W.C. Lee, A.S.A. Salam, M.F. Ibrahim, A.A.A. Rahni, and A.Z. Mohamed, "Autonomous Industrial Tank Floor Inspection Robot", IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2015, pp. 473-475.

[2] M. Makky, Delviyanti, and I. Berd, "Development of Aerial Online Intelligent Plant Monitoring System for Oil Palm (Elaeis guineensis Jacq.) Performance to External Stimuli," International Journal of Advanced Sciences Engineering Information Technology,Vol.8, 2018, pp. 2.

[3] H. Mansor, A. H. Adom, and N.A. Rahim, "Wireless Communication for Mobile Robots Using Commercial System," International Journal on Advanced Science Engineering Information Technology, Vol.2, 2012, pp. 1.

[4] R. Ramli, M.M. Yunus, and N.M. Ishak, "Robotic Teaching for Malaysian Gifted Enrichment Program," Procedia - Social and Behavioral Sciences, Vol.15, 2011, pp. 2528-2532.

[5] N. F. A. Zainal, R. Din, N. A. A. Majid, M. F. Nasrudin, and A.H.A. Rahman, "Primary and Secondary School Students Perspective on Kolb-based STEM Module and Robotic Prototype," International Journal of Advanced Sciences and Advanced Science Engineering Information Technology,Vol.8, 2018, pp. 2.

[6] Aliff, M., Yusof, M. I., Sani, N. S., & Zainal, A, Development of Fire Fighting Robot (QRob). International Journal of Advanced Computer Science and Applications (IJACSA), 10(1), 2019.

[7] Sani, N. S., Shamsuddin, I. I. S., Sahran, S., Rahman, A. H. A and Muzaffar, E. N, Redefining selection of features and classification algorithms for room occupancy detection, International Journal on Advanced Science, Engineering and Information Technology, 8(4-2), pp. 1486-1493, 2018.

[8] Sani, N.S., Rahman, M.A., Bakar, A.A., Sahran, S. and Sarim, H.M, "Machine learning approach for bottom 40 percent households (B40) poverty classification," International Journal on Advanced Science, Engineering and Information Technology, 8(4-2), pp.1698-1705, 2018.

[9] Holliday, J. D., Sani, N., & Willett, P. Calculation of substructural analysis weights using a genetic algorithm. Journal of chemical information and modeling, 55(2), pp. 214-221, 2015.

[10] Holliday, J. D., N. Sani, and P. Willett, Ligand-based virtual screening using a genetic algorithm with data fusion, Match: Communications in Mathematical and in Computer Chemistry, 80, pp. 623-638, 2018.

[11] J.-Y. Lee, Y.A. Song, J.Y. Jung, H.J. Kim, B.R. Kim, H.-K. Do, and J.-Y. Lim, "Nurses' Needs for Care Robots in Integrated Nursing Care Services," Journal of Advanced Nursing, Vol.74, 2018, pp. 2094-2105.

[12] K.M. Goher, N. Mansouri, and S.O. Fadlallah, "Assessment of Personal Care and Medical Robots from Older Adults' Perspective," Robotics and Biomimetics,Vol.4, 2017, pp. 5.

[13] C. Bayon, R. Raya, S. L. Lara, O. Ramírez, I.S. J, and E. Rocon, "Robotic Therapies for Children with Cerebral Palsy: a Systematic Review," Translational Biomedicine, Vol.7, 2016, pp. 44.

[14] B.S. Peters, P.R. Armijo, C. Krause, S.A. Choudhury, and D. Oleynikov, "Review of Emerging Surgical Robotic Technology." Surgical Endoscopy, Vol.32, 2018, pp. 1636-1655.

[15] H. Zheng, R. Davies, H. Zhou, J. Hammerton, J. Mawson Sue, M. Ware Patricia, D. Black Norman, C. Eccleston, H. Hu, T. Stone, A. Mountain Gail, and D. Harris Nigel, "SMART project: Application of emerging information and communication technology to home-based rehabilitation for stroke patients", International Journal on Disability and Human Development. 2006. p. 271.

[16] N. Aliman, R. Ramli, and S.M. Haris, "Design and development of Lower Limb Exoskeletons: A survey." Vol.95, 2017, pp. 102-116.

[17] M. M. Said, J. Yunas, B. Bais, A. A. Hamzah, and B.Y. Majlis, "The Design, Fabrication, and Testing of an Electromagnetic Micropump with a Matrix-Patterned Magnetic Polymer Composite Actuator Membrane." Micromachines, Vol.9, 2018, pp. 13.

[18] M. Aliff, S. Dohta, and T. Akagi, "Simple Trajectory Control Method of Robot Arm Using Flexible Pneumatic Cylinders." Journal of Robotics and Mechatronics.Vol.27, 2015, pp. 698-705.

[19] S. Dohta, T. Akagi, M. Aliff, and A. Ando. "Development and Control of Simple-structured Flexible Mechanisms using Flexible Pneumatic Cylinders", IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pp. 888-893, 2013.

[20] M. Aliff, S. Dohta, T. Akagi, and H. Li, "Development of a Simple-structured Pneumatic Robot Arm and its Control Using Low-cost Embedded Controller." Procedia Engineering,Vol.41, 2012, pp. 134-142.

[21] M. Aliff, S. Dohta, and T. Akagi, "Control and analysis of robot arm using flexible pneumatic cylinder." Mechanical Engineering Journal.Vol.1, 2014, pp. DR0051-DR0051.

[22] M. Aliff, S. Dohta, and T. Akagi. "Trajectory control of simple-structured flexible mechanism using flexible pneumatic cylinders", Proceedings of the 2013 IEEE/SICE International Symposium on System Integration, 2013, p. 19-24.

[23] M. Aliff, S. Dohta, T. Akagi, and T. Morimoto, "Control of Flexible Pneumatic Robot Arm Using Master Device with Pneumatic Brake Mechanism." JFPS International Journal of Fluid Power System.Vol.8, 2014, pp. 38-43.

[24] M. Aliff, S. Dohta, and T. Akagi, "Control and Analysis of Simple-structured Robot Arm using Flexible Pneumatic Cylinders." International Journal of Advanced and Applied Sciences, Vol.4, 2017, pp. 151-157.

Attractiveness Analysis of Quiz Games

Tara Khairiyah Md Zali[1], Nor Samsiah Sani[2], Abdul Hadi Abd Rahman[3], Mohd Aliff[4]
Faculty of Information Science and Technology (FTSM), Universiti Kebangsaan Malaysia (UKM)[1]
Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia (UKM)[2, 3]
Instrumentation and Control Engineering, Malaysian Institute of Industrial Technology
Universiti Kuala Lumpur, Johor Bahru, Johor[4]

***Abstract*—Quiz games are played on platforms such as television game shows, radio game shows, and recently, on mobile apps. In this study, HQ Trivia and SongPop 2 were chosen as the benchmark. Each game data have been collected for the analysis and the game refinement measure was employed for the assessment that focuses on different elimination tournament system for each sample. The results show that games such as HQ Trivia, which applies single-round elimination tournament, has a lower value of game refinement, in which the game is highly skillfull. Meanwhile, games that apply a round-robin system, such as SongPop 2 have a higher value of game refinement, in which the game is very stochastic. SongPop 2 and HQ Trivia both have more than 5 million downloads in Google Play Store. It is concluded that different types of quiz games which apply different kinds of tournament style have different game refinement value.**

***Keywords*—Quiz games; game refinement theory; attractiveness**

I. Introduction

The Merriam-Webster dictionary defines knowledge as the fact or condition of knowing something with familiarity gained through experience or association [1]. Knowledge is usually acquired through experience or education by perceiving, discovering, or learning new things. Ricci et al. investigated the effects a gaming approach on knowledge and retention in military trainees, which shows that participants assigned to game condition scored significantly higher on a retention test than those assigned to the text condition [2]. This indicates that people receive information better within game condition compared to the usual paper-based question-and-answer form (test).

Nowadays, there are many educational games available on various subjects, ranging from historical mythology to science and technology [3, 4]. Through games such as Age of Mythology and Age of Empires, one can learn about the mythological figures and popular culture superheroes and their connection to history and society [5]. In terms of learning science, one can simply browse the Science Kids website, which offers experimental science and technology games for kids to learn science in interactive ways. Furthermore, machine learning [6–10], used within learning analytics, provides new insights into education processes. Currently, it is used to develop quiz games in order to make the educational processes in schools and universities more efficient for students and teachers.

One of the factors available in the quiz games that attract people into playing it applies to the game itself. Most of the quizzes offer a variety of categories of quiz questions, and the players may or may not choose a category of their liking. Some of them make the quiz game much more challenging by limiting the time, treating that as a goal to answer the questions as fast as possible. Another main factor is the offer of prize money to the winners, which gameplays mainly apply. For example, the television game shows attract people to participate and watch the players play by correctly answering the questions in the hope of winning the game and returning home with a huge sum of money [11]. It is believed that to maximize the entertainment factor of the game, game designers have to find a comfortable setting for the quiz game [12]. Hence, factors that attract the quiz game need to be identified, applied, and changed consecutively.

A previous study has used quiz games to identify the point of popularity that attracts people to play it. Quiz games have been played for so long – ever since the radio started broadcasting to the public. The questions that may be asked in the quiz games vary. For instance, there are quiz games specially made to test players' knowledge of music or television series. With respect to the root question of "Why are people playing quiz games for a long time?" the trend gained a lot of ground in the '70s with the original Jeopardy! Daytime game shows premiered in 1964 [13]. By applying game refinement theory to quiz games by using an appropriate game model, this study focuses on two main research goals: 1) to find the reason why quiz games have been popular for such a long time and 2) to identify comfortable settings of quiz games.

The rest of this paper is organised as follows. Section II addresses the background of study related to quiz games. Section III explains the game refinement theory for attractive analysis. The quiz analysis is presented in Section IV. This paper is concluded in Section V.

II. Background of the Study

A. Historical Review

The period on which the term quiz was created is unidentifiable. The American Heritage Dictionary mentions that in 1782, quiz was apparently an unrelated slang word that meant an odd person or an eccentric person; this definition was originally derived from the term quizzical [14].

Additionally, the dictionary suggests that it may come from the English dialect “quiset”, which means “to question”. In this case, it may originate as a question and refer to inquisitiveness [15]. Based on the Oxford English Dictionary [16], the term quiz means “to question or interrogate”, which may originate from a statement recorded in the year 1843, “She comes back an’ quiesed us”.

Quiz games have been played ever since the existence of radio shows. In America, the earliest radio quiz show was Information Please, which was aired on NBC from 17th May 1938 to 22nd April 1951. The title of the show originated from the contemporary phrase used to request information from telephone operators. Then, it was called “information”, but now, it is called directory assistance. The series was moderated by Clifton Fadiman.

B. Overview of Quiz Games

Mainly, the goal of quiz games is usually to answer all the questions correctly. Nowadays, there are many types of quiz games. Some have the intention to win the game by answering the questions within the time limit, while some are played to beat the opponent’s score. These kinds of quiz gameplay have evolved without anyone knowing where and when it all started. Some quizzes offer various kinds of categories, in which players can pick the category of questions they would like to answer, and some are randomly picked questions that usually deal with general knowledge.

Basically, the goal of any quiz game is to make its players win by correctly answering all the questions given. Some gameplays are conducted by counting the participant’s highest score mark and considering the event of the participant defeating other players’ highest marks. To achieve this, a player has to win by answering a lot of questions correctly. For example, a player begins by registering his/her minor details to the game. Then, the player starts to play by answering beginner’s level questions. With every question answered, the rank of the player in the game increases. As the rank increases, the player is then challenged with much more challenging questions than the ones before. As for quiz games that offer prize money, they usually use the single-elimination (SE) tournament system, in which the players must win by answering all the questions correctly within the time limit. If the players answer even one question wrong, they are automatically disqualified from the game.

III. Research Methodology

To undertake these challenges, this paper focuses on two parts of quiz games, which are the gameplay of quiz games, and the questioning part of quiz games. For each part, this study attempts to figure out the reasonable game progress model to derive an appropriate measure of game refinement. The data have been collected using a variety of methods. For example, some data were collected through playing the game itself in order to identify the gameplay of quiz games, while some data were obtained from reliable sources on the internet.

This project has implemented a game refinement theory as defined by Sutiono Purwarianti and Iida [17]–the “game progress” is twofold. One is game speed or scoring rate, while the other is the game information progress which focuses on the game outcome. In quiz games, the scoring rate is calculated by two factors: 1) number of questions correctly answered and 2) the time taken to answer the question. Thus, the game speed is given by the average amount of questions divided by the number of total mistakes. In some quiz games, the total score may solely depend on the total number of correctly answered questions instead of depending on the time taken to answer the questions.

Now, considering a model of game information progress, the game information progress itself indicates the certainty of the result of the game in a certain time. Having full information regarding the game information progress, let G be the number of total mistakes and T be the average number of questions. As for game information progress, for example, after the game, the game progress will be given as a linear function of time t with $0 \le t \le T$ and $x(t) \le G$ as shown in Eq. (1) [18].

$$x(t) = \frac{G}{T} t \qquad (1)$$

However, the game information progress given by Eq. 1 is usually not known during the in-game period. This is because of the presence of uncertainty during the game until it ends, which is called balanced game or seesaw game. Therefore, the game information progress should not be linear but rather exponential. Hence, a realistic model of game information progress is given in Eq. (2).

$$x(t) = G\left(\frac{t}{T}\right)^{n} \qquad (2)$$

Here, *n* stands for a constant parameter that is given based on the perspective of the observer in the game. By deriving Eq. (2) twice, the acceleration of game information progress is obtained. Eq. (3) presents the final equation when solving it a t = T.

$$x^{n}(T) = \frac{Gn(n-1)}{T^{n}} t^{n-2} = \frac{G}{T^{2}} n(n-1) \qquad (3)$$

In this study, it has been assumed that the game information progress in any type of game occurs in human brains. The physics of information in the brain is not known yet, and it is likely that the acceleration of information progress was related to the forces and laws of physics. Hence, it is reasonably expected that the larger the value of G/T^2, the game becomes more exciting due to the uncertainty of the game outcome. Thus, we have used Eq. (4) as a game refinement measure for the game under consideration. It is called R value in short.

$$R = \frac{\sqrt{G}}{T} \qquad (4)$$

Here, the gap between board games and sports games has been considered by deriving a formula to calculate the game information progress of board games. Let *B* be an average branching factor (number of possible options) and *D* the game length (depth of whole game tree). Table I shows the measurement of game refinement for board games (i.e., Chess, Go, and Mahjong).

A round in board games can be illustrated as a decision tree. At each depth of the game tree, one choses a move, and the game progresses. Fig. 1 illustrates one level of game tree. The distance *d* can be found using a simple Pythagoras Theorem, as shown in Eq. (5).

$$d = \sqrt{\Delta l^2 + 1} \tag{5}$$

Assuming the approximate value of horizontal difference between nodes is B/2, the substitution results in Eq. (6).

$$d = \sqrt{\left(\frac{B}{2}\right)^2 + 1} \tag{6}$$

The game progress for one game is the total level of game tree times *d*. For the meantime, it is not considered because the value *d* is assumed to be much smaller compared to *B*. The game length is normalised by the average game length *D*; then, the game progress $x(t)$ is given by Eq. (7).

$$x(t) = \frac{t}{D} \cdot d = \frac{t}{D}\sqrt{\left(\frac{B}{2}\right)^2} = \frac{Bt}{2D} \tag{7}$$

TABLE I. MEASURES OF GAME REFINEMENT FOR BOARD GAMES

Game	**B**	**D**	**R**
Chess	35	80	0.074
Go	250	208	0.076
Mahjong	10.36	49.36	0.078

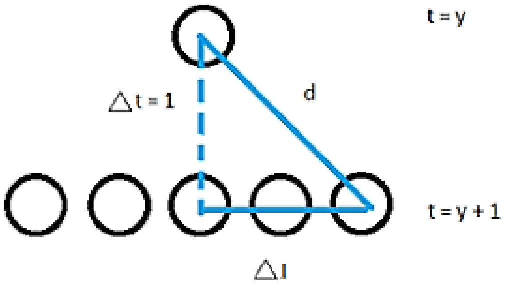


Fig. 1. One Level of Game Tree Illustration.

IV. ANALYSIS OF QUIZ GAMES

A list of questions and answers is the core of a quiz game. To analyse a quiz, it is necessary to focus on this part first. Our first approach was to collect data by searching the information through the game's official website. If the information in the website was deemed not enough, we experimented by playing the game. Moreover, by using this approach, we have collected data from a much more reliable source. Finally, the analysis was conducted to answer our main research purposes.

A. Quiz Gameplay

In this project, five main features were selected to determine quiz attractiveness, which consists of multiple-choice questions (MCQ), time limit scoring system, high-score list, and types of tournament. The details of each feature are as follows:

1) Multiple-Choice Question (MCQ): This consists of several possible answers, from which the correct one must be selected [5]. The multiple-choice format was found to yield more reliability and validity in a shorter amount of test-taking time as compared with short-answer tests [19]. Mainly, quiz games use the MCQ option, where some games offer around 2–4 answer options, from which the players have to pick only one correct answer.

2) Time limit: Many sophisticated board games and popular time limit sports games have a similar value of game refinement [17]. By setting time limit to the game, it introduces challenge into a game in the form of timed response. Players are needed to complete every question within the assigned time limit. Failure to do so may lead to the player losing the game. Time limit is effective for being challenging because it introduces an explicit goal that is not trivial for players to achieve if the game is not properly calibrated [20].

3) Scoring system: One of the most direct methods of motivating players is by assigning points for each and every correct answer during the game. Using points increases players' motivation by providing a clear connection with the effort shown in the game [21]. Furthermore, a score summary following each game provides players with performance feedback as well as facilitates progress assessment on beating the goals of the game.

4) High-Score list: Another method for motivating players to play is by using the high-score list, which shows the names and scores of the players who have achieved the highest scores. The score needed to be beaten by players are shown in order to identify the goal of beating the high score. In a quiz game played by category, the high score is given specifically. By doing this, players become more motivated in answering all of the questions correctly so that they can beat the high score.

5) Types of tournament: There are various ways to run a tournament, but there are about two formats that are popular within the quiz game, which are SE tournament and Round Robin tournament.

B. Data Collection

One possible way to collect the data of quiz games is by experimenting with the games themselves. As most official websites of quiz games do not provide adequate information regarding the games themselves, the only way left is by experimenting with them. Therefore, in this study, two quiz games were simulated by simplifying the factors in the game. The detail of the games is as follows.

1) SongPop 2: SongPop 2 is a music trivia game that was released in July 2015 by FreshPlanet. It is a free-to-play app with in-app purchases. The game is similar to the popular American television game show "Name That Tune", which tests player's knowledge about songs. SongPop 2 features over 100,000 songs and 1,000 curated playlists. Currently, it has more than 5 million downloads in Google Play Store.

Here, players first pick the type of tournament they want. There are three types of mode that players can choose to play: the single-player mode, one-to-one mode, and multiplayers mode. They first pick the music category that they want to play. Then, they listen to the song being played and choose the correct answer from the four options provided in the multiple-choice questions. Some questions ask for the title of a song, and some ask about the singer of a song. The players have only ten seconds to answer each question for a total of 10 questions per round. The faster the players pick the correct answer, the higher their score become. SongPop 2 practices the Round-Robin (RR) elimination tournament, where players can afford to make a number of wrong answers without being eliminated from the game. Table II shows the game mode and game details of SongPop2.

In this version of SongPop 2, players can compete in party mode against hundreds of players in daily multiplayers tournament, where players compete to win badges. Additionally, they can play a single-player mode in which a player competes against the computer. This is the improved version of the game as compared to the earlier version of Songpop 1, where there was only the option of competing with only one opponent. With respect to the SongPop scoring formula, it is awarded based on time and how many consecutive answers players have in their streak, which is completely dependent on the previous questions.

2) HQ trivia: HQ Trivia was released in August 26, 2017 on iOS and later for Android on December 31, 2017. It is developed by Vine creators, Rus Yusupov and Colin Kroll. It is a free-to-play quiz game with in-app purchases that offers prize money to the players who manage to correctly answer a series of questions with increasing difficulty. The app is inspired by a live game show that is aired at 9 pm (the US time). There are around 300,000 players per game in HQ Trivia with 2 million players playing HQ Trivia. Currently, it has more than 5 million downloads in Google Play Store.

Players have ten seconds to answer each multiple-choice question for a total of twelve questions. If there is more than one player who has managed to correctly answer the questions, the prize money is split equally among them. Each question has three possible answers. The players have 10 seconds to answer each question. HQ Trivia game practices SE tournament, in which the players who wrongly answer or do not manage to answer in the limited time are automatically eliminated from the match. Fig. 2 shows the probability of the number of players left with each wrongly answered question.

TABLE II. GAME DETAILS OF SONGPOP2

Game Mode	**Practice Mode**	**One-to-one**	**Multiplayer**
Opponent	Computer	1	4
No of Questions	5	5	10
No of Errors	5	5	10

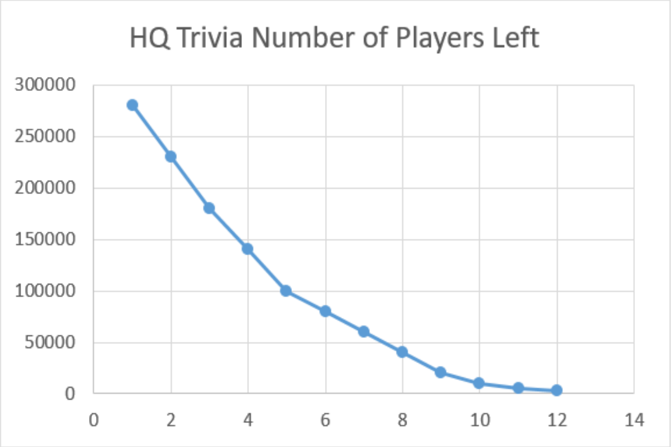


Fig. 2. Probability of Number of Players Left Per Each Question.

C. Discussion

Throughout the analysis of quiz game apps between HQ Trivia and SongPop 2, two quiz game aspects were found which were round system aspect and tournament style. Fig. 3 shows the quiz game aspects for the measures of game refinement for quiz games. For the round system aspect, a time-limit approach has been used with the game refinement measure variable for Eq. (8).

$$GR = \frac{\sqrt{G}}{T} \tag{8}$$

The variable G has been identified as the number of error that can be made, and T is identified as the total number of questions per round. The value *n* in the tournament style aspect refers to the number of participants' entry.

Table III shows the comparison between HQ Trivia app and SongPop 2, which identifies each game's refinement measure value. The round system aspect for quiz games uses the time-limit approach. As for HQ Trivia that applies SE tournament type, players answer a total of 12 questions, and they are automatically eliminated if they answer a question wrong even once. The GR-value of 0.08, which is within the game sophistication zone value $0.07 \le GR \le 0.08$, for HQ Trivia indicates that the game is highly competitive and entertaining at the same time. For SongPop 2 that applies the RR tournament type, the GR-value is $0.3 \le GR \le 0.5$, which is higher than the game sophistication value. Therefore, it can be deduced that the game depends heavily on the players' luck.

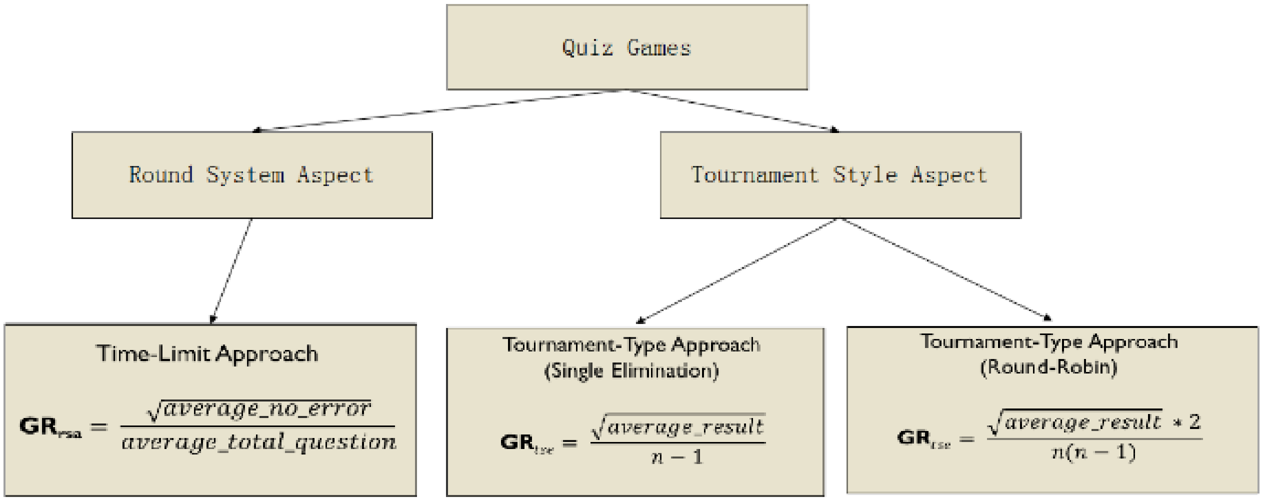

Quiz Games
Round System Aspect
Tournament Style Aspect
Time-Limit Approach
$\mathbf{GR}_{rsa} = \frac{\sqrt{average\_no\_error}}{average\_total\_question}$
Tournament-Type Approach (Single Elimination)
$\mathbf{GR}_{tse} = \frac{\sqrt{average\_result}}{n-1}$
Tournament-Type Approach (Round-Robin)
$\mathbf{GR}_{tse} = \frac{\sqrt{average\_result} * 2}{n(n-1)}$

Fig. 3. Quiz Games Aspects.

TABLE III. COMPARISON BETWEEN HQ TRIVIA APP AND SONGPOP2

	HQ TRIVIA	SongPop 2		
Tournament Type	Single Elimination	Round Robin		
		Practice Mode	One-to-One	Multi-player
Total Questions	12	5	5	10
Total Mistakes	1	5	5	10
No. Of Entries	120 000	2	2	5
Possible Results	14	1	1	1
GR_{RSA}	0.083	0.447	0.447	0.316
GR_{TSE}	0.00003	1	1	0.1

As for tournament style aspect, it was divided between the SE type approach and the RR type approach. The SE type quiz eliminates the player once they make an error in answering the question applied in HQ Trivia. The GR-value of the tournament-style aspect for HQ Trivia is lower than the game sophistication zone value, which is 0.00003, in which the minimum value of zone sophistication value is 0.07. Furthermore, HQ Trivia is highly dependable on the player's skills due to the increasing difficulty level of the questions.

As for the RR approach, players have to answer every question without getting eliminated from the game. This kind of approach is applied in SongPop 2. The games with RR approach apply a scoring system to identify the winners. SongPop 2 players are rewarded based on how fast they answer the question and the total bonus marks for consecutive correct answers. The GR-value of the tournament-style aspect for SongPop 2 is quite high, 0.1~1.0. The maximum value of zone sophistication is 0.08. Additionally, SongPop 2 is highly stochastic or unpredictable and depends heavily on players' luck.

For quiz games using the SE tournament setting, its round aspect's GR-value is lower than games that apply the RR tournament setting. A SE tournament quiz such as HQ Trivia has a value of 0.08, which implies that it has both the balance of competitiveness and entertainment. In contrast, SongPop 2 with RR tournament setting recorded a value of 0.3~0.5, which implies that it was highly stochastic and depends heavily on chances. As for the tournament style aspect, quiz games that use the SE tournament setting have a GR-value was lower than the game sophistication zone value recorded at 0.00003. Furthermore, games that apply the SE tournament setting tend to be highly dependable on the players' skills. Moreover, quiz games that apply the RR tournament setting have a value of 0.1~1.0, which implies that they are highly stochastic and depend heavily on chances. Thus, we have concluded that different types of quiz games that apply different kinds of tournament styles have different game refinement values.

V. CONCLUSION

Quiz games have been popular ever since the radio started broadcasting them; currently, they are being played on television and smartphones. The game refinement measure for quiz games has been calculated for two types of quizzes that has different settings of tournament. This study presents an attractiveness analysis for quiz games that can be used to help refine the development of future quiz games. With deeper knowledge on the refinement value of quiz games, an additional number of quiz games can be used to generalise the game refinement value. Apart from that, an observation can be made to keep track of the game data such as number of the players and number of winners of the game in order to get reliable data.

ACKNOWLEDGMENT

The authors would like to thank the Universiti Kebangsaan Malaysia and Ministry of Education Malaysia (MOE) under the Research University Grants (project codes: GGPM-2017-039 and FRGS/1/2018/ICT02/UKM/02/6) for funding and supporting this research.

REFERENCES

[1] Merriam-Webster, Inc, 2004. Merriam-Webster Dictionary Online. Merriam-Webster's collegiate dictionary. Retrieved August, 6, p. 2018, 2004.

[2] K. E. Ricci, E. Salas, and J. A. Cannon-Bowers, "Do computer-based games facilitate knowledge acquisition and retention?" Military Psychology, vol. 8(4), pp. 295–307, 1996.

[3] D. Siegle, "Technology: Learning can be fun and games," Gifted Child Today, vol. 38(3), pp. 192–197, 2015.

[4] K. Osman and N. A. Bakar, "Educational computer games for Malaysian classrooms: Issues and challenges," Asian Social Science, vol. 8(11), pp. 75, 2012.

[5] J. P. Gee, "What video games have to teach us about learning and literacy", Computers in Entertainment (CIE), vol. 1(1), p. 20, 2003.

[6] M. Aliff, M. I. Yusof, N. S. Sani, and A. Zainal, "Development of fire fighting robot (QRob)," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10(1), 2019.

[7] N. S. Sani, I. I. S. Shamsuddin, S. Sahran, A. H. A. Rahman, and E. N. Muzaffar, "Redefining selection of features and classification algorithms for room occupancy detection," International Journal on Advanced Science, Engineering and Information Technology, vol. 8(4–2), pp. 1486–1493, 2018.

[8] N. S. Sani, M. A. Rahman, A. A. Bakar, S. Sahran, and H. M. Sarim, "Machine learning approach for bottom 40 percent households (B40) poverty classification," International Journal on Advanced Science, Engineering and Information Technology, 8(4-2), pp.1698-1705, 2018.

[9] J. D. Holliday, N. Sani, and P. Willett, "Calculation of substructural analysis weights using a genetic algorithm," Journal of Chemical Information and Modeling, vol. 55(2), pp. 214–221, 2015.

[10] J. D. Holliday, N. Sani, and P. Willett, "Ligand-based virtual screening using a genetic algorithm with data fusion," Match: Communications in Mathematical and in Computer Chemistry, vol. 80, pp. 623–638, 2018.

[11] J. Haigh, "TV game shows," in: Mathematics in Everyday Life. Cham. :Springer, 2016, pp 113–131.

[12] M. Watson and L. Bozgeyikli, "Introduction to game theory via an interactive gameplay experience," in Companion Publication of the 2019 on Designing Interactive Systems Conference, 2019, pp. 319–323.

[13] S. Jaya, GQ: Why Do People Love Trivia So Much?. https://www.gq.com/story/why-do-people-love-trivia-so-much.. Retrieved August, 20, p. 2018, 2018.

[14] M. Berube, ed., "The American Heritage Dictionary" Second College Edition. Houghton Mifflin. Retrieved August, 6, p. 2018, 1985.

[15] K. E. Ricci, E. Salas, and J. A. Cannon-Bowers, "Do computer-based games facilitate knowledge acquisition and retention?" Military Psychology, vol. 8(4), pp. 295–307, 1996.

[16] J. Simpson and E. S. Weiner, Oxford English Dictionary Online. Oxford: Clarendon Press. Retrieved August, 7, p. 2018, 1989.

[17] A. P. Sutiono, A. Purwarianti, and H. Iida, "A mathematical model of game refinement," International Conference on Intelligent Technologies for Interactive Entertainment, Cham.: Springer, pp. 148–15, 2014.

[18] S. Xiong and H. Iida, "Attractiveness of real time strategy games," Systems and Informatics (ICSAI), 2014 2nd International Conference, IEEE, pp. 271–276, 2014.

[19] D. R. Bacon, "Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context," Journal of Marketing Education, vol. 25(1), pp. 31–36, 2003.

[20] N. Nossal and H. Iida, "Game refinement theory and its application to score limit games," Games Media Entertainment (GEM) IEEE, pp. 1–3, 2014.

[21] L. Von Ahn and L. Dabbish, "Designing games with a purpose," Communications of the ACM, vol. 51(8), pp. 58–67, 2008.

A Survey: Agent-based Software Technology Under the Eyes of Cyber Security, Security Controls, Attacks and Challenges

Bandar Alluhaybi[1], Mohamad Shady Alrahhal[2], Ahmed Alzhrani[3], Vijey Thayananthan[4]
King Abdulaziz University (KAU), Jeddah, Saudi Arabia

***Abstract*—Recently, agent-based software technology has received wide attention by the research community due to its valuable benefits, such as reducing the load on networks and providing an efficient solution for the transmission challenge problem. However, the major concern in building agent-based systems is related to the security of agents. In this paper, we explore the techniques used to build controls that guarantee both the protection of agents against malicious destination machines and the protection of destination machines against malicious agents. In addition, statistical-based analyses are employed to evaluate the level of maturity of the protection techniques to preserve the protection goals (the code and data, state, and itinerary of the agent), with and without the threat of attacks. Challenges regarding the security of agents are presented and highlighted by seven research questions related to satisfying cyber security requirements, protecting the visiting agent and the visited host machine from each other, providing robustness against advanced attacks that target protection goals, quantifying the security in agent-based systems, and providing features of self-protection and self-communication to the agent itself.**

Keywords—Agent; attack; cyber; security; requirement; maturity; protection goals

I. Introduction

One of the most important software technologies that is used to manage and perform tasks over the Internet is agent-based software technology (ABST). A software agent is defined as an independent program that runs run on behalf of a network user [1, 2, 3]. ABST has been involved in many research fields, varying from network management tasks to information management ones [4, 5]. The power of the ABST is inspired by its valuable properties. The properties of this technology can be summarized as follows [3, 6]:

- Mobility, which is a unique property of this technology. It means that the agent can move from one machine to another machine, performing a specific mission there, and then it must come back to the original machine with the results. In other words, the agent is goal-driven.
- Adaptability, which means platform independent. In other words, this property enables the agent to be executed on different machines regardless of the operating system used.
- Transparency and accountability, which explains that the software agent runs on behalf its owner, and the owner of the agent can ask the agent about its current location and about what has been accomplished.
- Ruggedness, which refers to the capability of the agent to run on either low or high resources and to interpret different data formats.
- Self-start or proactive means that the time of starting a mission, the time of finishing a mission, and the time of delivering results are features that are based on the knowledge of the agent and have no relationship to the owner of the agent.

Thus, the agent is not restricted by the machine where it is written, but it has the ability of moving among machines via a network [7]. This action is called migration, as shown in Fig. 1.

In Fig. 1, different machines are connected via a network. The owner of the agent creates it at a machine called the home machine (HM). Then, the mobile agent can migrate to other machines called destination machines (DMs, where each DM has its own operating system (OS) as well as its own hardware (HW) specifications. A uniform agent manager, which is middleware (MW), is installed at the HM and at each DM. Concordia [8], Java Agent Development Framework (JADE) [9], and Agelets [10] are examples of agent managers. Fig. 2 shows a code example written by Concordia to illustrate the migration process of the mobile agent.

In the example above, the owner creates the agent, then creates an itinerary to move it to a DM called "dbserver", then back to an HM called "workstation". The agentsCodebase and relatedClasses specify the objects containing the methods and data necessary to complete the mission. More specifically, an itinerary is created, and when the agent ready to migrate, it prepares a list of its intended destinations. The itinerary of the agent is used by the Concordia server to determine the network destination of the agent. With each method included in the itinerary (i.e., "queryDatabase" and "reportResults"), the local Concordia server will move the agent and its objects to the machine specified in the nest itinerary entry. When the itinerary is exhausted, the trip of the agent is finished. Therefore, the itinerary caused the agent to move to "dbserver" and execute the "queryDatabase" method, then to move back to "workstation" and execute the "reportResults" method. The last line of the code illustrates an additional argument to the "launchAgent" method, which causes the codebase and the "QueryResults" class definitions to travel with the agent.

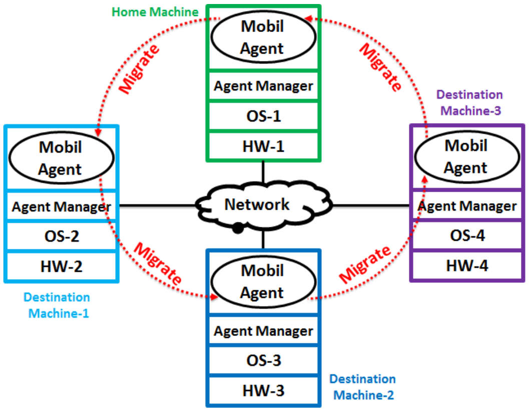
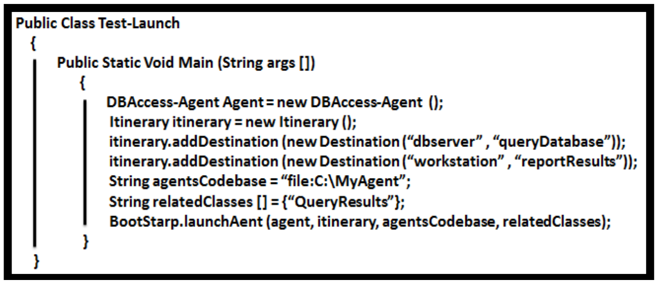


Fig. 1. Migration of Mobile Agent.



```
Public Class Test-Launch
 {
    Public Static Void Main (String args [])
      {
        DBAccess-Agent Agent = new DBAccess-Agent ();
         Itinerary itinerary = new Itinerary ();
         itinerary.addDestination (new Destination (“dbserver” , “queryDatabase”));
         itinerary.addDestination (new Destination (“workstation” , “reportResults”));
         String agentsCodebase = “file:C:\MyAgent”;
         String relatedClasses [] = {“QueryResults”};
         BootStarp.launchAent (agent, itinerary, agentsCodebase, relatedClasses);
      }
 }
```

Fig. 2. Concordia Code for Mobile Agent Migration.

From the description above, four parts of the mobile agent are travelling during the migration. They are:

- The code of the mobile agent.
- The data, which are manipulated by the code.
- The itinerary information, which includes the HM and the DM.
- The state, which describes data controlled by the CPU and OS and includes the results of the executed mission.

Statement of the problem however, the mobile agents can be targets for attackers, where any one of the parts (or all of them) listed above can be the victim. This in turn can shoot the functionalities of any agent-based system in the heart. Specifically, passive attacks (such as eavesdropping and repudiation attacks) or active attacks (such as alternation and replay attacks) can be applied to the agents involved in the system. In passive attacks, the information carried by the mobile agent can be stolen to be misused later for malicious purposes [11]; meanwhile, in active attacks, the carried information is obtained and modified during the migration for the purpose of performing malicious actions [12]. The two previous kinds of attacks can be performed by an external attacker (i.e., located between the HM and the DM), but do not address the scenario in which the DM itself is the attacker. In this case, the danger may reach severe levels because the DM has full control of the execution of the hosted agent. On other hand, the mobile agent may itself be the attacker, with the ability of launching or performing poisonous pieces of codes against the DM. As a result, ensuring the security of the mobile agents as well as protecting the DMs against malicious agents is a pressing issue.

In this survey, we review the different techniques proposed previously to ensure the security in the software agent research field as well as the potential cyber-attacks. The contribution of this paper is as follows:

- We provide a statistical model called the maturity model to evaluate the protection mechanisms in agent-based systems. The maturity model relies on the protection goals, which represent the main parts of the mobile agent.
- We employ the maturity model for both evaluating the protection mechanisms under threats of different attacks and ranking the attacks according to their danger.
- We summarize the challenges of security of the agent's research field by seven research questions.

The rest of the paper is structured as follows. In Section II, we highlight the importance of agent-based software technology. Section III provides the cyber security requirements in agent-based systems. A classification of security techniques is presented in Section IV. Section V discusses achieving the cyber security requirements. In Section VI, the protection goals, attacks, and maturity model-based analyses are discussed. Section VII provides a strategy to evaluate an agent-based system with the security metrics that can be used. The challenges and the corresponding research questions are presented in Section VIII. Finally, we conclude the paper in Section IX.

II. Importance of ABST in Distributed Systems

Before the birth of ABST, many client-server-based technologies have been used for developing distributed systems, such as message passing (MP) [13], remote procedure call (RPC) [14], and Code on Demand (CoD) [15]. Under the interaction term between the client and the server, AGST overcomes the previous technologies, as shown in Fig. 3.

Fig. 3 shows that in MP, RPC, and CoD technologies, if a user wants to send (n) requests to the server, the network channel is occupied by n sizes of the requests. After processing the requests at the server side, the network channel will be occupied by n sizes of the responses. Formally, let $size_R$ denote the size of the request and $size_P$ denote the size of the response. BW denotes the band width of the network channel, and NT denotes the network traffic. Then,

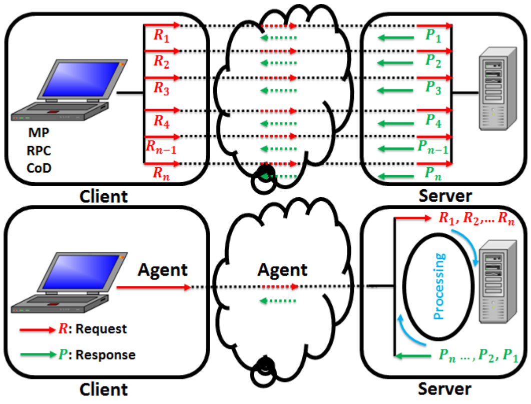


Fig. 3. ABST vs. other Technologies.

$$\mathrm{NT}_{\mathrm{out-coming}}\ \% = \frac{\mathrm{BW}}{\mathrm{n} \times \mathrm{size}_{\mathrm{R}}} \quad (1)$$

$$\mathrm{NT}_{\mathrm{in-coming}}\ \% = \frac{\mathrm{BW}}{\mathrm{n} \times \mathrm{size}_{\mathrm{P}}} \quad (2)$$

In the worst case, the interaction between the client and the server will be at a high level, which requires interleaving among the out-coming requests and the in-coming responses within the network channel. Thus,

$$\mathrm{NT}\ \% = \frac{\mathrm{BW}}{\mathrm{n} \times (\mathrm{size}_{\mathrm{R}} + \mathrm{size}_{\mathrm{P}})} \quad (3)$$

Compared to MP, RPC, and CoD technologies, the network channel is occupied by only the size of the migrated agent ($\mathrm{size}_{\mathrm{agent}}$), which is very small compared to ($\mathrm{n} \times \mathrm{size}_{\mathrm{R}}$). This, in turn, contributes to reduce the network traffic efficiently. After processing the requests at the server side (i.e., executing the mission of the agent), the agent migrates back to the client, carrying the responses included in the state part. It is worth mentioning that there is no worst case in the ABST. However, during the migration back, the four parts (i.e., state, code, data, and itinerary information) are included within the agent. This will increase the network traffic slightly.

More formally, let ($\mathrm{size}_{\mathrm{st}}$), ($\mathrm{size}_{\mathrm{co}}$), ($\mathrm{size}_{\mathrm{da}}$), and ($\mathrm{size}_{\mathrm{it}}$) refer to the sizes of the state, code, data, and itinerary information respectively. Then, the size of the agent during migration ($\mathrm{size}_{\mathrm{agent\_m}}$) is defined as:

$$\mathrm{size}_{\mathrm{agent\_m}} = \mathrm{size}_{\mathrm{co}} + \mathrm{size}_{\mathrm{da}} + \mathrm{size}_{\mathrm{it}} \quad (4)$$

It is obvious that:

$$\mathrm{size}_{\mathrm{agent\_m}} < \mathrm{n} \times \mathrm{size}_{\mathrm{R}} \quad (5)$$

During the agent's migration back and because of the results' size included in the state, the size of its state is:

$$\mathrm{size}_{\mathrm{st}} = \mathrm{n} \times \mathrm{size}_{\mathrm{P}} \quad (6)$$

Thus, the size of the agent during migration back ($\mathrm{size}_{\mathrm{agent\_mb}}$) is updated to be:

$$\mathrm{size}_{\mathrm{agent_{mb}}} = (\mathrm{size}_{\mathrm{co}} + \mathrm{size}_{\mathrm{da}} + \mathrm{size}_{\mathrm{it}}) + (\mathrm{n} \times \mathrm{size}_{\mathrm{P}}) \quad (7)$$

Because no interleaving exists when using agents, the size of the agent that migrates back is less than the sum of the sizes of the requests and responses of the worst case in other technologies. This is represented by the following formula:

$$\mathrm{size}_{\mathrm{agent\_mb}} < \mathrm{n} \times (\mathrm{size}_{\mathrm{R}} + \mathrm{size}_{\mathrm{P}}) \quad (8)$$

Thus, the NT % based on ABST is less than the NT % based on other technologies.

Moreover, under the first aspect of the scalability quality attribute (i.e., increasing the number of users), the ABST also overcomes other technologies. Let $\mathrm{M}_{\mathrm{user}}$ denote the number of users that are using a system, where each user sends n requests, each one of size $\mathrm{size}_{\mathrm{R}}$. Thus, the size of total number of sent requests ($\mathrm{size}_{\mathrm{T\_R}}$) is defined as:

$$\mathrm{size}_{\mathrm{T\_R}} = \mathrm{M}_{\mathrm{user}} \times \mathrm{n} \times \mathrm{size}_{\mathrm{R}} \quad (9)$$

Compared to MP, RPC, and CoD technologies, each user creates an agent of size $\mathrm{size}_{\mathrm{agent\_m}}$ in the matched agent-based system. Consequently, the size of the total sent agents ($\mathrm{size}_{\mathrm{T\_Ag}}$) is:

$$\mathrm{size}_{\mathrm{T\_Ag}} = \mathrm{M}_{\mathrm{user}} \times \mathrm{size}_{\mathrm{agent\_m}} \quad (10)$$

When increasing the number of users (i.e., $\mathrm{M}_{\mathrm{user}} = 1000$, 2000, …, 10,000 users), it is obvious that:

$$\mathrm{size}_{\mathrm{T\_Ag}} \ll \mathrm{size}_{\mathrm{T\_R}} \quad (11)$$

Furthermore, under the second aspect of the scalability quality attribute (i.e., increasing the size of the manipulated data), the ABST is efficient. Let ($\mathrm{size}_{\mathrm{mp\_da}}$) refer to the size of the manipulated data at the server side. When increasing the size of the manipulated data, for example, such that ($\mathrm{k} \times \mathrm{size}_{\mathrm{mp\_da}}$), where (k = 2, 4, 6, …, 10), the performance will not dramatically deteriorate. This is quite true when dealing with Big Data (BD) sizes [16, 17]. That lack of deterioration is because the mobile agents migrate to the machines where the BD is located, processing it there, and then returning back with results of manipulation only. This provides an efficient solution to what is called the transmission challenge, which occurs because small sizes of codes (i.e., agents) migrate via the network channel to end tasks, rather than transmitting huge sizes of BD to the manipulating machines [18, 19, 20].

Since the time is tightly coupled with the transmission, ABST can overcome the network latency, especially when manipulating health data and multimedia [21, 22]. Moreover, under access latency and tuning time terms [23], the ABST outperforms other technologies. Access latency refers the time elapsed between the moment when a request is issued and the moment when it is satisfied. Let (T_{AL}, $\mathrm{T}_{\mathrm{comp}}$, $\mathrm{T}_{\mathrm{sending}}$, $\mathrm{T}_{\mathrm{receiving}}$) denote the access latency, computation time, sending time, and receiving time respectively. Then, access latency is defined as:

$$\mathrm{T}_{\mathrm{AL}} = \mathrm{T}_{\mathrm{comp}} + \mathrm{T}_{\mathrm{sending}} + \mathrm{T}_{\mathrm{receiving}} \quad (12)$$

Suppose that the $\mathrm{T}_{\mathrm{comp}}$ is the same in both the ABST and other technologies. Because NT % based on ABST is less than NT % based on other technologies, both $\mathrm{T}_{\mathrm{sending}}$ and $\mathrm{T}_{\mathrm{receiving}}$ are short, which in turn leads to shorter access latency in the ABST. Tuning time (T_{TU}) is defined as the time a machine of a client stays active to receive the requested data. Since the T_{AL} is short in the ABST, the T_{TU} is also short compared to other technologies. This is quite true when the client uses his/her smart phone as a machine to send the requests and to receive the corresponding results, contributing to power consumption savings [24].

Table I highlights the characteristics of the ABST compared to other technologies.

Other benefits of agents, such as executing dynamically, asynchronously, and autonomously, are discussed in the work [25], where the authors provided seven good reasons for using the ABST.

TABLE. I. CHARACTERISTICS OF THE AGENT-BASED SOFTWARE TECHNOLOGY (ABST)

Term \ Technology		ABST	MP	RPC	CoD
Network traffic		Low	High	High	High
Scalability in big data	***Increasing NO. of users***	Efficient	Inefficient	Inefficient	Inefficient
	Increasing size of data	Efficient	Inefficient	Inefficient	Inefficient
Transmission challenge		Solved	Un-solved	Un-solved	Un-solved
Network latency		Low	High	High	High
Manipulation	***Access latency***	Short	Long	Long	Long
	Tuning Time	Short	Long	Long	Long

An additional feature is added to the ABST when comparing it to other technologies, such as component-based software technology (CBST) and web service-based software technology (WSBST). This feature is related to architecture building. In the CBST and WSBST, the architecture of the proposed distributed system is built during the design time. Meanwhile, the architecture is built during the run time in the ABST. Moreover, the ABST adopts the three types of architectures (i.e., sequential, parallel, and hybrid architectures). Actually, Fig. 1 above represents the sequential architecture, where the agent created at the HM visits a series of DMs in a sequential manner. Finally, the agent migrates back to the HM. Fig. 4 and 5 illustrate the parallel, and hybrid architectures, respectively.

In Fig. 4, the owner creates three different agents at the HM. Then, the agents are migrated in parallel to the corresponding DMs. In detail, the first agent migrates to DM-1 to perform its own task. The same scenario is followed by the second and third agent, where they migrate to DM-2 and DM-3 respectively. Finally, each agent migrates back to the HM with the results once its mission is finished.

Fig. 5 illustrates a hybrid agent-based architecture, where two different agents are created at the HM. The two agents start their itinerary in parallel, where the first agent visits DM-1 and DM-2, and then migrates back to the HM in a sequential manner. The second agent behaves the same, but its itinerary contains DM-3, DM-4, and DH-5.

Due to the benefits of the ABST explained above, it is involved in building a wide spectrum of distributed systems. Resource management in cloud computing [26], fault tolerance [27], distributed network performance management [28], security testing in web-based applications [29], and privacy protection in location-based services [30] are agent-based distributed systems, where agents play a significant role in performing the functionalities of these systems. However, again, the security of mobile agents at the interface remains a critical issue.

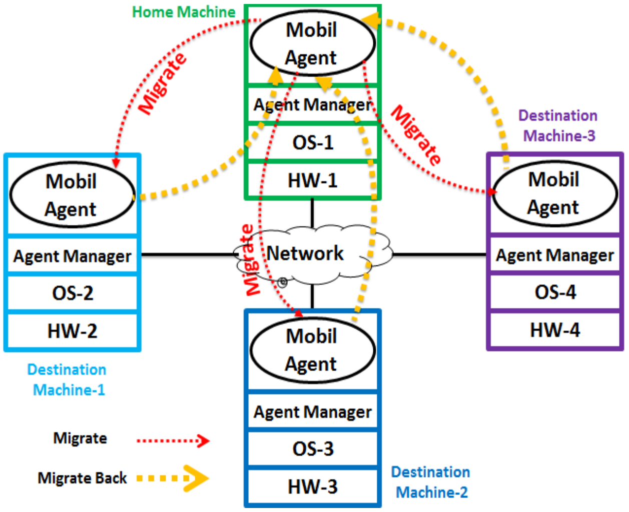


Fig. 4. Parallel Agent-based Architecture.

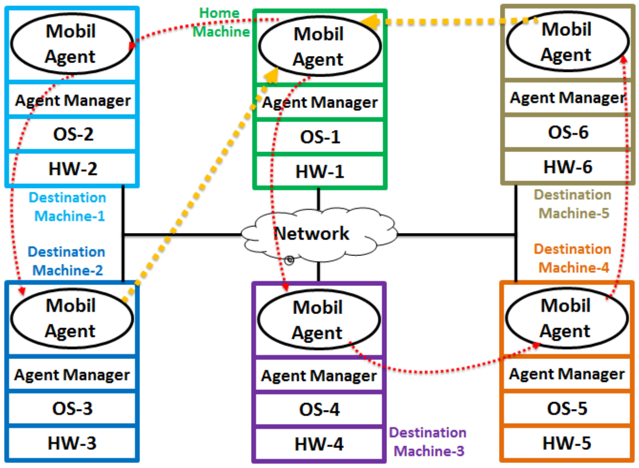


Fig. 5. Hybrid Agent-based Architecture.

III. CYBER SECURITY REQUIREMENTS IN THE LIFECYCLE OF MOBILE AGENT

This section explains the stages of the lifecycle of mobile agents and defines the cyber security requirements (CSR) needed to safely end the assigned mission.

A. Lifecycle of the Mobile Agent

There are three main stages involved in the lifecycle of the mobile agent, which are creation, migration, and termination, as shown in Fig. 6.

In the creation stage, the mobile agent is created (i.e., is written by the owner using a specific agent manager) with its itinerary as well as the mission that should be performed. This stage is conducted at the HM. In the migration stage, the mobile agent follows the path of the itinerary, visiting one or more DMs. After completing the itinerary, the agent is terminated. If the mobile agent safely returns to the HM, the termination is performed by the owner of the agent. Otherwise, it is killed or blocked by a visited DM (i.e., it is attacked). In other words, the danger to the mobile agent starts at the moment of leaving the HM, where it can be attacked during moving among DMs or by any of the visited DMs.

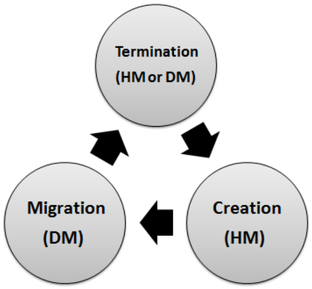


Fig. 6. Lifecycle of the Mobile Agent.

B. *Cyber Security Requirements*

If we want to represent cyber security, we can represent it as an umbrella under which two main aspects are located, which are: security and privacy. Ensuring security means establishing a secure communication between the sender and receiver to safely exchange messages, where cryptography is the core of the techniques used in this aspect. Meanwhile, privacy means protecting sensitive data against misuse by attackers. To highlight the importance of the privacy aspect, we can consider location-based services (LBS), for example, where the user sends a query asking for the nearest hospitals. In LBS-enabled applications, the user is forced to reveal sensitive data, such as the real location and the queried Point of Interest (PoI), which in turn reflects personal aspects in his/her realistic life, such as a religious or health state. Such sensitive data can be exploited and misused later by attackers for blackmail or mugging. Indeed, the authors of [31] and [32] provided surveys on the techniques used to protect the privacy of the user, where the dummy-based technique [32, 33] is considered a powerful approach for this purpose.

In distributed systems, the key security requirements are represented by the CIA triads (i.e., confidentiality, integrity, and availability); meanwhile, the key privacy requirements are represented by the TLI triads [34, 35] (i.e., tractability, linkability, and identifiability). Fig. 7 illustrates the security and privacy triads under the cyber security umbrella.

The CIA represents traditional security requirements for designing and implementing any distributed system. However, using the ABST demands additional security requirements, which are the Six A's (i.e., anonymity, accountability, authentication, authorization, accounting, and assurance) as well as non-repudiation and verification. Table II below describes the CSR needed in an agent-based distributed system.

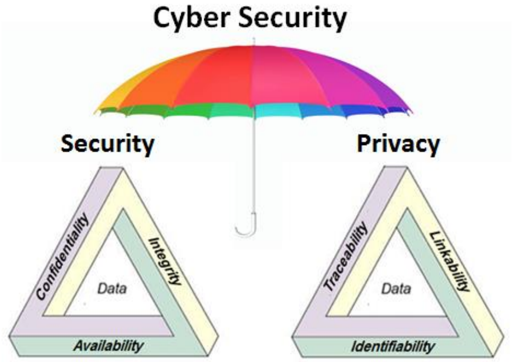


Fig. 7. Security Triads and Privacy Triads.

TABLE. II. CYBER SECURITY REQUIREMENTS

Term **Aspect**	**Abbr./Name**	**Description**
Security	***Traditional security requirements***	
	C (confidentiality)	The information carried by the mobile agent must be kept secret and only authorized parties can access it.
	I (Integrity)	Guarding the carried information against improper modification or destruction.
	A (Availability)	The assurance that the carried data are accessible when needed by authorized parities, including users and DMs.
	Six A's	
	An (Anonymity)	Achieving load balancing between keeping the actions of the agent private and auditing the agent when utilizing/logging the resources of the DMs.
	Ac (Accountability)	All actions that are performed on a DM should be traceable to the agent who committed them (i.e., logs should be kept, archived, and secured).
	Au (Authentication)	The positive identification of both the agent seeking access to a current DM and the carried information from a previous machine in an itinerary before execution of the mission on the current DM.
	Ar (Authorization)	The act of granting the agent actual access to information resources of the DM, where the level of access may change based on the agent's defined access level.
	At (Accounting)	The logging of access and usage of the DM's resources. In other words, keeping track the agent who accesses what resource, when, and for how long.
	As (Assurance)	The controls used to develop confidence that security measures are working as intended. Auditing, monitoring, testing, and reporting are the foundations of assurance.
	Additional security requirements	
	Non-R (Repudiation)	The agent platform that sends the information to an agent owner or other DM cannot deny that he is the owner of the specific information and agent.
	Ve (Verification)	Only the authenticated mobile agent is permitted access into the DM, and the code of the migrated agent from the HM is verified before execution.
Privacy	T (Tractability)	The ability to verify the history, location, or application of an agent by means of documented recorded identification.
	L (Linkability)	The attacker can sufficiently distinguish whether two or more agents are related or not within the system.
	I (Identifiability)	The attacker can sufficiently identify the entities within the system.

Both security and privacy must be considered as top quality attributes in designing agent-based distributed systems. Consequently, the CSR mentioned above should be satisfied over all stages of the agent's lifecycle, to ensure that the system is perfect under the cyber security term.

IV. Classification of Security Techniques in Mobile Agents

There are two main classes of approaches proposed in the research field on the security of mobile agents, which are the approaches that secure the agent platform (i.e., DM) against malicious mobile agents and the approaches that secure the mobile agents against malicious platforms. Each class has its own techniques, as shown in Fig. 8.

There are two main classes of approaches proposed in the research field on the security of mobile agents, which are the approaches that secure the agent platform (i.e., DM) against malicious mobile agents and the approaches that secure the mobile agents against malicious platforms. Each class has its own techniques, as shown in Fig. 8.

A. *First Class: Protecting the Agent Platform*

In this class, the attacker is the malicious mobile agent that visits a DM, and the victim is the DM platform. Many techniques have been proposed to protect the DM platform as described below.

1) Sandboxing: This is a software technique that depends on the principle of isolation of the execution of the suspected code in a virtual space under tight restrictions. Relying on the sandboxing technique, the authors in [36] proposed a mechanism that enforces the mobile agent to follow a fixed security policy for execution its code. This mechanism succeeds in preventing the mobile agent from (1) interacting with the local file system; (2) accessing the system properties; and (3) opening a network connection. Under this technique, an enhanced approach was proposed by Noordende et al. in [37]. The authors focused on the restrictions that deal with memory to prevent the unauthorized access by the poisonous code.

However, the major drawback of the sandboxing technique is that it consumes a long execution time (due to the strict restrictions) even if the mobile agent's code is legal.

2) Code signing: This technique targets ensuring the integrity of the code that is executed on the DM platform. It tunes with both the one-way hash functions and the digital signature concepts to ensure that no modification is done on the code. Therefore, this technique assumes that the creator of the code is trusted. The authors of the work [38] provide shining proof about the resistance of this technique, where it is used in ActiveX controls and Java applets. An enhanced verification-based approach is introduced by Malik et al. [39]. Their key idea depends on using white and black lists of entities, where a security manager checks the incoming code. If it is coming from a trusted entity (i.e., included within the white list), the code is then granted full permissions to be executed. Otherwise, it will be frozen.

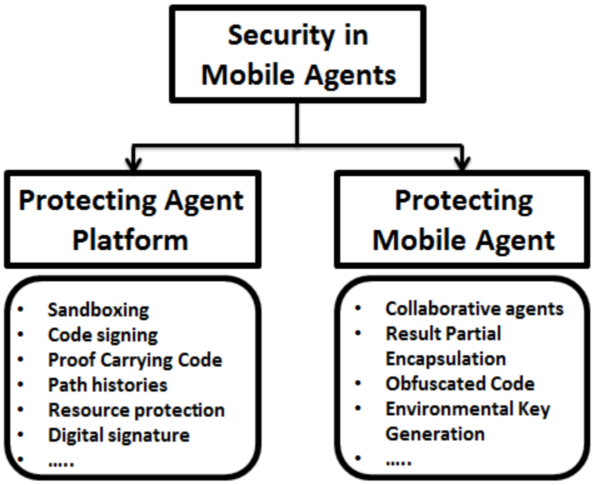


Fig. 8. Classification of Security Approaches in Mobile Agents.

However, the main drawback of this technique is that it requires the continuous update of both white and black lists, which is a large obstacle in light of the changing dynamic nature of entities as time progresses. In addition, it is computationally costly due to using hash functions in addition to encryption and decryption [40].

3) Proof Carrying Code (PCC): In this technique, the creator of the code marks the code (i.e., generates a proof attached to the original code), so that any modification that occurred will be detected and the code is not allowed to be executed. Compared to the code signing technique, the PCC is better regarding time and computation costs. The reason behind this is that the PCC does not require cryptography for the digital signature. In the work [41], the authors proposed a foundational proof-carrying code, in which the code is verified with the smallest possible set of axioms, using the simplest possible verifier and the smallest possible runtime system. An enhanced PCC-based technique is presented in [42], where the major concern is allowing dynamic access to the platform of DM, with a tolerance of the strict proof representation.

However, a sharp criticism was directed at this technique in [43], where the proof generation is the main problem with PCC, as well as the automation of this process.

4) Path histories: This technique embraces the principle of ascertaining the level of trust of the visited DMs' platforms during the life cycle of the agent. Therefore, the mobile agent is forced to maintain an authenticable trajectory of the previously visited DMs. Relying on this technique, the authors of the works [44] and [45] proposed two approaches that grant the mobile agent suitable privileges that match the corresponding trust levels.

However, the problem of this technique is explained in [46]. The cost of checking the trust level increases when the number of visited DM involved in the path history is increased. Moreover, it is complex to predict the trust level of the DMs visited in the future that are included in the path history in advance.

5) Resource protection: This technique employs the fundamentals of authentication to allow only legal agents to access the resources of the DM. Thus, the platform of the agent at the DM is protected. An authentication-based proxy is proposed in [12], where the mobile agent is not allowed to access resources unless it reveals its identity using the public and private keys. Another approach is presented with this technique in [47]. The authors' idea was inspired from the real-world scenario, in which the mobile agent can prove its identity by providing a passport and visa, which carries information that describes the credibility of the agent.

However, the limitation of this technique is related to the proxy overhead computation. In addition, the identity (i.e., the passport) may be stolen or impersonated.

6) Digital signature: It is a common technique used in secure communication networks that satisfies confidentiality and integrity. It is similar to the code signing technique, but the difference is applying a digital signature on the mobile agent itself instead of the carried code. A digital signature-based approach supported by a checkpoint mechanism is provided in [48]. The objective of the checkpoint mechanism is to guarantee the validity of the mobile agent using fragmentation and defragmentation methods. Based on both the digital signature and verifying method, another approach is proposed in [49]. In this work, the authors mix the code signing technique with the digital signature technique. The code of the agent is signed by the creator, and the code is executed at the DM after being verified by the owner of the agent.

However, supporting the digital signature-based technique by fragmentation and verification leads to a trade-off between the strength of the proposed approach and the computation cost.

7) Policy-based model: In this technique, predefined diagnosis methods are applied to the mobile agent once reaching the DM. Based on the results of the diagnosis, the agent is allowed or not allowed to execute. A malicious content scanning-based approach is presented in [50]. The scanner provides an alarm to the DM if any suspected content exists. An immune system is proposed in [51]. Actually, the work [51] is considered a development of the work [50], where performance was the axis of the enhancement. The key idea is to employ the pipelining concept in scanning, predicting, and extracting the malicious piece of code.

However, although the performance is enhanced, the process of scanning and discovering the malicious content is still costly due to the different u of the mobile agents' executions.

8) State appraisal. This technique tunes with the state carried by the mobile agent in a pure programming way. In depth, a maximum set of safe permissions that the agent could request from the DM is encapsulated within a state appraisal function, depending on the agent's current state. Based on this technique, a state appraisal function is proposed in [52] to ensure the security of the DM. The agent calls the state appraisal function to retrieve the permissions of the current visited DM and does not violate them. Then, when the mobile agent leaves the current DM moving to the next one, the state appraisal function is called again. Thus, the previous state, which represents the input of the mission that should be performed at the next DM, is ensured. In this way, the next DM guarantees that the state was not modified, and consequently, the arriving agent is not malicious. Similar to [52], the authors of [53] rely on the state appraisal function, but the difference is supporting the function by an authentication mechanism between the sender of the mobile agent from the current DM and the receiver of the mobile agent at the next DM.

However, the major issue in this technique is the difficulty of formulating and adopting the mobile agent with the security permissions of each visited DM.

9) Machine learning: This technique employs data mining concepts to protect the visited DM, depending on a classification data mining task. Recently, a supervised machine learning classifier was proposed in [54] by Pallavi et al. The authors used a data set that contains 80 mobile agents (half of them are malicious, and the remaining are non-malicious). Then, the features of all agents are extracted to determine the behaviors of the agents. Finally, using the extracted features, a decision tree-based algorithm is applied to the data set to make the execution decision related to a mobile agent. Depending on the data mining classification task, another approach is introduced in [55]. The same strategy used in [54] is used in [55], but the difference is that the authors used the K-nearest neighbor algorithm to build the classifier instead of the decision tree-based algorithm.

However, the main obstacle encountered with this technique is the excessive expense of building a good knowledge data base with a large number of agents. In addition, using different classifiers leads to different results.

Second Class: Protecting the Mobile Agent

In this class, the attacker is the visited DM and the victim is the mobile agent. Many techniques have been proposed to protect the mobile agent against the DM, as described below.

10) Collaborative agents: The principle of this technique relies on sharing secret information about the sensitive tasks between two cooperating agents, so that the DM cannot steal and tamper with the trajectory of the itinerary. Depending on this technique, a secure communication protocol between the agents is proposed in [56]. This protocol establishes an authenticated communication channel between the cooperating agents to share the content of the itinerary of the first agent with the second agent. The content of the itinerary adjusts the triads of visited DMs (i.e., the previous/last visited DM, the current DM, and the future DM). The second cooperating agent takes the responsibility for manipulating any inconsistency that may occur, such as the current DM sending the agent to the wrong future DM or generating an alarm about receiving the agent from a wrong source. Madkour et al. proposed another cooperative approach to protect mobile agents against malicious DMs in [57]. The key idea is to create an assistant agent, called the shadow that follows the original one. If the original agent is attacked by the DM, it informs the shadow, kills itself, and the shadow in turn sends an acknowledgement to inform the owner of the original agent about the attack. The shadow then becomes the original agent, and the owner of the agent creates a new agent to be a new shadow.

However, the gap of this technique is the cost of configuration and establishing the authenticated communication channel for each migration.

11) Result Partial Encapsulation (RPE): This technique is designed to detect any changes that might occur regarding the results of an executed mission at a DM by a mobile agent. To end this, the results are encapsulated so that a verification step is performed later at the HM to provide proof that no change was performed by an attack. This technique is applied on the agent's code to provide confidentiality using encryption based on the secret key [58]. The key idea is to have a list of secret keys stored within the mobile agent, used for encryption, such that each key is related to a specific DM. In the current DM, the agent uses the corresponding secret key to generate message authentication code (MAC). Then, encapsulating the MAC with a message that represents the results of the mission execution generates partial result authentication code (PRAC). Based on the RPF technique, the authors of [59] proposed an approach to ensure both confidentiality and integrity of the results using a digital signature. This approach is called sliding encryption, which aims at decreasing both the time processing and the required storage by encrypting a small amount of data. The sliding encryption approach is developed so that it can be adopted in certain applications where storage space is valuable, such as smartcards.

However, the main drawback of this technique is ensuring future integrity, where the next DM can obtain the secret key of the previous DM to modify its generated results.

12) Obfuscated Code: In this technique, the mobile agent travels through series of DMs that have different trust levels. To ensure that no DM is able to extract sensitive data hidden in the code (such as the secret key or credit card number), the behavior of the mobile code is protected. The core protection performs some obfuscating transformations on the code before actual execution, so that the code cannot be understood by the malicious DM. Based on the obfuscation code technique, Hohl et al. [60] proposed the black box security approach to preserve the behavior of the code. They obfuscated the data structure used within the code without modifying the code itself. Another approach is provided in the context of this technique in the work [61]. The difference here is the way of modifying the code, where the control flow in the code is modified without affecting the computing part of the code.

However, the main challenge of this technique is adopting it to suit different applications, where behaviors can extensively change from one application to another.

13) Environmental Key Generation: This technique relies on the principle that states "the execution is not allowed unless some environmental conditions are satisfied at the DM". In the work [62], the authors defined the environmental conditions as matching a specific search string. When this condition is true, an activation key is performed to allow the execution. The activation key function is hidden within a file system. Similar to [62], the authors of the work [63] used the same condition, but the difference is that the activation key is included within the content of an email.

However, the limitation associated with this technique is that the DM may act maliciously after the condition is satisfied and the activation key is performed. Moreover, the key activation may be a virus file to be executed at the DM. Therefore, the DM tries to not allow execution even if the condition is satisfied.

14) Execution tracing: This technique targets discovering malicious modifications that may be performed by DM on the mobile agent code, state, and execution flow. The scenario followed by this technique consists of three steps, which are (1) a DM that receives the agent and agrees to execute it and produce an associated trace during the agent's execution; (2) a message is attached by DM to the mobile agent, containing information about the unique identifier of the message, the identity of the sender, the timestamp, the fingerprint of the trace, and the final state carried by the agent; and (3) the HM (i.e., the owner of the agent) asks the DM to provide the previous message (the trace) to validate it by comparing it with a fingerprint generated by the agent. In [64], a detailed protocol for message exchanging is provided to adjust the previous three steps in a mathematical manner. An enhancement is achieved on the previous protocol by Tan et al. in [65]. The key enhancement is assigning the mission of the trace validation to a trusted third party (TTP) instead of the owner of the agent. The TTP here is called validation or verification server.

However, the execution tracing technique suffers from the potential malicious collaboration between the validation server and a DM.

15) Watermarking: Originally, the watermarking term refers to the process of embedding a watermark within an information entity, such as image, audio, video, or text files for copyright protection purposes. The authors of [66] exploit the watermarking technique to detect an attack that aims at modifying the results of the mobile agent's mission execution. Consequently, the results are watermarked, and if a DM attacked them, the embedded watermark is damaged or destroyed. To detect the occurrence of the attack, the watermark is extracted at the HM and compared with the original one. The work [66] is developed by the same authors in [67] to be adopted with various kinds of watermarks. Therefore, during execution, the agent can employ any kind of available information as a watermark, such as dummy data, input data, intermediate variable values, or data originating from communications.

However, the watermarking technique has a critical gap, which is that the embedded watermark can be destroyed by a compression attack. Compression attacks can be performed by any external attacker (i.e., not by the DM). Thus, the DM is considered malicious while it is not.

16) Co-signing: This technique relies on hiring an external trusted party to co-sign the migration of the agent. In [68], the preceding DM is considered the external party, which acts as an observer by taking the responsibility of co-signing the mobile agent. Actually, the work [68] is proposed to give mobile agents resistance against multiple colluded DMs that target poisoning the results of execution. Another approach is presented in [69] based on the co-signing technique. The key idea is that after producing the results, the DMs encapsulate them with the information of the mission carried by the mobile agent. Then, the entire encapsulated package is encrypted and sent to the next DM at the same time. When the mobile agent reaches the next DM, a comparison is performed between the generated results and the mission information to discover any attack that may have occurred.

However, time consumption, network overhead, and robustness against Denial of Service (DoS) attacks are considered the main challenges in this technique, especially when the mobile agent carries a time-sensitive task.

17) Separation of privileges: The essence of this technique is managing the agent-based system by separating the tasks and assigning them to some major agents. The goal of this technique is to minimize the capabilities of the malicious DMs to attack the visiting agents. In [70], three agents control the system, which are controller agent (CA), worker agent (WA), and itinerary register agent (IRA). The CA is responsible for storing and manipulating the core data. The WA is responsible for storing and manipulating functions that have less importance than the previous one. The IRA is responsible for storing the addresses of the visited DMs, and the time at which the execution is performed on each DM. The authors of [71] followed the same strategy as that presented in [70]. The difference is that the privileges are supported by roles.

However, the process of extracting the privileges, separating them, and supporting them with accurate roles that control different cases that may be involved leads to increasing the complexity of the agent-based system.

18) Fragmentation-based encryption: This technique aims at enhancing the performance, where only the sensitive data that may be exploited by a DM are first extracted. Then, these sensitive data are encrypted. Finally, the encrypted sensitive data are randomized so that only the agent knows the process of backing the correct order. In [72], the bytes of the agent's code are scanned, and then the sensitive parts are encrypted and inserted within predefined arrays. When execution at the DM occurs, the agent uses the same randomization key (i.e., the seed) to retrieve the correct ordering of all code bytes. Similar to [72], the protocol proposed in [73] depends on a fragmentation technique. The difference is that the extraction, encryption, and randomization stages are performed by a TTP.

However, despite the performance enhancement achieved through encrypting only the sensitive data of the agent's code, the process of generating the seed of the randomization algorithm, applying the algorithm, and reordering the randomized code may lead (in some cases) to exceeding the time needed for encrypting all of the agent's code.

V. Achieving Cyber Security Requirements

Table III (a, b) below compares the approaches discussed in the previous section in terms of satisfying the CSR.

Drawn conclusions from Table III (a, b), the following observations can be made:

1) Among the six A's, anonymity and accountability security requirements are not achieved in all approaches related to protecting the agent platform. Actually, there is a clear and strong trade-off between these two security requirements, as it is obvious from their concepts. Anonymity and accountability security requirements are critical for the second class (i.e., protecting the mobile agent).

2) Linkability and identifiability privacy requirements are not achieved in all approaches related to protecting the agent platform. Therefore, the attacker (malicious agent) has the ability of revealing some entities of the system within the DMs. For example, in [50] and [51], the policy file used for protecting the agent's platform is known by the visiting agent. The policy file may reflect the sensitivity level of the system installed on the DM where the mobile agent is executed.

3) Among CIA, the availability security requirement is not achieved in all approaches related to protecting the mobile agent class, which is normal. Actually, the availability security requirement is critical for protecting the agent platform class, where not achieving it means that the DMs suffer from the DoS attack.

4) All of the six A's are critical and should be achieved in any approach related to protecting the mobile agent class.

5) The non-repudiation and verification additional security requirements are not achieved in all approaches related to protecting the mobile agent class. Actually, these additional security requirements are critical for protecting the agent platform class, so that the code of the agent is verified before execution and the HM cannot deny creating the mobile agent and sending it to the DMs.

6) All privacy requirements (TLI) were completely ignored and were not addressed in all approaches related to protecting the mobile agent class.

7) The authorization, accounting, and assurance security requirements are mandatory necessities and should be satisfied by both classes.

TABLE. III. (A) SYMBOLS

Symbol	**C**	**I**	**A**	**An**	**Ac**	**Au**	**Ar**
Based on	Confidentiality	Integrity	Availability	Anonymity	Accountability	Authentication	Authorization
Symbol	**At**	**As**	**Non-R**	**Ve**	**T**	**L**	**I**
Based on	Accountability	Assurance	Non-Repudiation	Verification	Tractability	Linkability	Identifiability

(B) SATISFYING THE CYBER SECURITY REQUIREMENTS

	CSR	**Security aspect**											**Privacy aspect**		
Class	**Tech**	***CIA***			***Six A's***						***Add. SR***		***TLI***		
		C	I	A	An	Ac	Au	Ar	At	As	Non-R	Ve	T	L	I
Protecting agent platform	***Sandboxing***														
	[36]	√	√	√	×	×	√	×	√	×	×	×	√	×	×
	[37]	√	√	√	×	×	×	×	×	×	×	×	√	×	×
	Code signing														
	[38]	√	×	×	×	×	√	×	×	×	×	×	√	×	×
	[39]	√	×	×	×	×	√	×	×	×	×	√	√	×	×
	PCC														
	[41]	√	×	×	×	×	×	×	×	×	×	√	√	×	×
	[42]	√	√	×	×	×	×	×	×	×	×	√	√	×	×
	Bath history														
	[44]	×	√	×	×	×	√	√	×	√	√	×	×	×	×
	[45]	×	√	×	×	×	√	√	×	√	√	×	×	×	×
	Resource protection														
	[12]	√	√	√	×	×	√	×	√	√	×	√	×	×	×
	[47]	√	√	√	×	×	√	×	√	√	×	√	×	×	×
	Digital signature														
	[48]	√	√	×	×	×	√	×	×	×	×	×	√	×	×
	[49]	√	√	×	×	×	√	×	×	×	×	×	√	×	×
	Policy-based model														
	[50]	√	√	√	×	×	×	×	√	×	√	√	×	×	×
	[51]	√	√	√	×	×	×	×	√	×	√	√	×	×	×
	State appraisal														
	[52]	√	×	×	×	×	×	√	×	×	×	×	√	×	×
	[53]	√	×	×	×	×	√	√	×	×	×	×	√	×	×
	Machine learning														
	[54]	√	√	√	×	×	×	×	×	×	×	×	×	×	×
	[55]	√	√	√	×	×	×	×	×	×	×	×	×	×	×
Protecting mobile agent	***Collaborative agents***														
	[56]	×	√	×	×	√	√	×	×	×	×	×	×	×	×
	[57]	×	√	×	×	√	√	×	×	×	×	×	×	×	×
	RPE														
	[58]	√	×	×	×	√	×	×	×	×	×	×	×	×	×
	[59]	√	√	×	×	√	√	×	×	×	×	×	×	×	×
	Obfuscated code														
	[60]	√	×	×	×	√	×	×	×	×	×	×	×	×	×
	[61]	√	×	×	×	√	×	×	×	×	×	×	×	×	×
	Environment key generation														
	[62]	√	×	×	×	×	√	√	√	√	×	×	×	×	×
	[63]	√	×	×	×	×	√	√	√	√	×	×	×	×	×
	Executing tracking														
	[64]	√	√	×	×	√	√	×	×	√	×	×	×	×	×
	[65]	√	√	×	×	√	√	×	×	√	×	×	×	×	×
	Watermarking														
	[66]	×	√	×	×	×	×	×	×	×	×	×	×	×	×
	[67]	×	√	×	×	×	×	×	×	×	×	×	×	×	×
	Co-signing														
	[68]	√	√	×	×	×	√	×	√	×	×	×	×	×	×
	[69]	√	√	×	×	×	√	×	√	×	×	×	×	×	×
	Separation of privileges														
	[70]	√	√	×	×	√	×	√	√	√	×	×	×	×	×
	[71]	√	√	×	×	√	×	√	√	√	×	×	×	×	×
	Fragmentation-based encryption														
	[72]	×	√	×	√	×	×	×	×	×	×	×	×	×	×
	[73]	√	√	×	×	×	√	×	×	×	×	×	×	×	×

TABLE. IV. SECURITY REQUIREMENTS ACCORDING TO THE CLASS OF SECURITY AGENTS

Class	Distinguished security requirements
protecting mobile agent	Anonymity, Accountability
protecting agent platform	Availability, Non-Repudiation, Verification
Security requirements needed in both classes	
Confidentiality, Integrity, Authentication, Authorization, Accounting, and Assurance	

Based on the conclusions drawn and represented by the points discussed above, we distinguish the security requirements that are individually needed for each of the two classes as well as those needed for both. Table IV separates the security requirements according to the classes. It is worth mentioning that all privacy aspects are needed for the two classes.

VI. PROTECTION GOALS AND ATTACKS

When a mobile agent migrates from an HM to a DM, it might be attacked by the DM. In this section, we define the protection goals that an approach aims to guarantee in protecting the mobile agent class. Then, we explore the potential attacks and measure their impacts on the protection goals based on a maturity-based model.

A. Protection Goals

As mentioned in the introduction section, the mobile agent consists of four main parts, which are the code, the data, the state, and the itinerary. Since the code of the mobile agent and the data are tightly coupled, we refer to them as the first protection goal. The second and third protection goals are the state and the itinerary respectively. To explain how these goals are attacked, we provide an example inspired from a smart city environment, as described below.

In smart cities, ensuring comprehensive safety is an important issue for saving people's lives. Smart warning systems (SWSs) can contribute to achieve this noble goal through alarming the decision makers to take the corresponding steps that ensure avoiding disasters [74, 75, 76]. Fig. 9 illustrates the general concept of SWS.

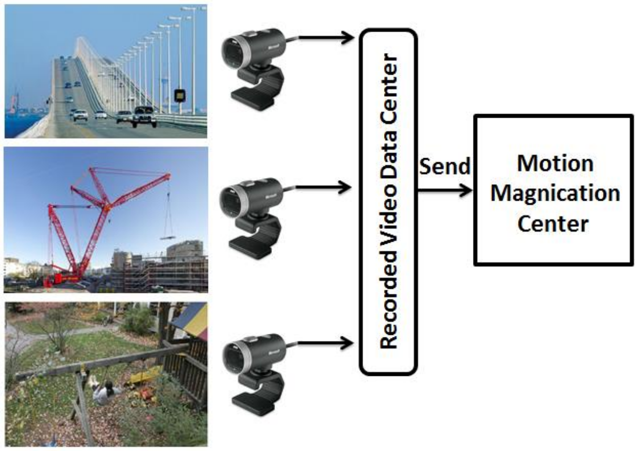


Fig. 9. Smart Warning System.

In Fig. 9, fixed cameras record the motion of different objects in smart cities. The motion magnification centre (MMC) processes the recorded video to enlarge the unseen, abnormal, and critical motions that may cause disasters. Color clustering-based and phase-based video motion processing techniques, which resemble a microscope that amplifies subtle motions in a video sequence allowing visualization of deformations that would otherwise be invisible, can be found in [77, 78].

When agent software technology is employed to build such an SWS, the architecture of the SWS is illustrated in Figure 10.

As shown in Fig. 10, a mobile agent is created at the decision-making centre (the HM), and this mobile agent then migrates to the recording video centre (the DM) to perform a magnification task on the recorded video. After performing the magnification task at the DM, the mobile agent migrates back to the HM carrying the results to be analysed at the decision-making centre for the purpose of avoiding disasters.

As a first scenario, the code and the data of the mobile agent as well as the code of the task can be attacked by the DM, so that a modification can skew the task intended to be performed. As a second scenario, the code and the data are not modified, where the task is performed correctly at the recording video centre, but the results of the task execution are modified. In other words, the mobile agent will carry the wrong results to be analysed at the decision-making centre. As a third and final scenario, when many DMs are involved in the itinerary, the itinerary information of the mobile agent can be changed so that the agent migrates to the wrong next DM. However, in the all three scenarios, the disaster may occur with a high probability.

To avoid any of dangerous scenarios mentioned above, the protection mechanisms should guarantee the protection of the three goals at the same time. In terms of goals' protection, Table V compares the approaches proposed to protect the mobile agent against the DM as an attacker.

As inferred from Table V, the majority of the approaches succeed in protecting the first goal (17 out of 18), while they fail in protecting the second and third goal (1 out of 18 and 2 out of 18, respectively).

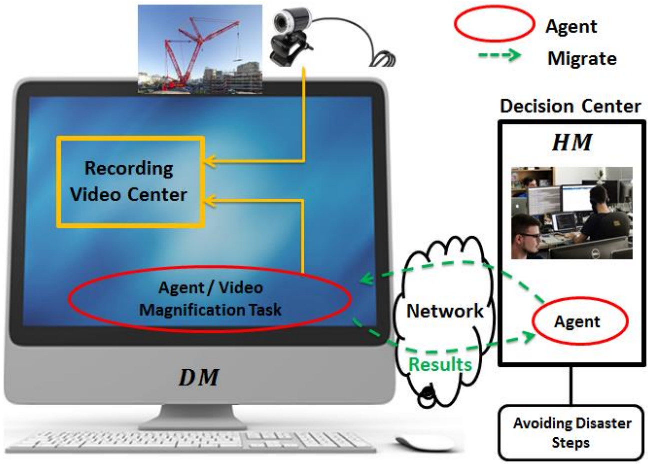


Fig. 10. Agent-based SWS Architecture.

TABLE. V. ACHIEVING PROTECTION GOALS IN PROTECTING MOBILE AGENT APPROACHES

Term		Protection Goals		
Category	**Tech**	***Code & Data***	***State***	***Itinerary***
Protecting mobile Agent	***Collaborative agents***			
	[56]	√	×	√
	[57]	×	×	√
	RPE			
	[58]	√	×	×
	[59]	√	×	×
	Obfuscated code			
	[60]	√	×	×
	[61]	√	√	×
	Environment Key Generation			
	[62]	√	×	×
	[63]	√	×	×
	Execution tracking			
	[64]	√	×	×
	[65]	√	×	×
	Watermarking			
	[66]	√	×	×
	[67]	√	×	×
	Co-signing			
	[68]	√	×	×
	[69]	√	×	×
	Separation of Privileges			
	[70]	√	×	×
	[71]	√	×	×
	Fragmentation based encryption			
	[72]	√	×	×
	[73]	√	×	×

B. Attacks

Exploitation of vulnerabilities is actually considered the spirit of the attacks. Therefore, there is a strong relation between the attacks and vulnerabilities. The protection mechanisms or measures address the control of the vulnerabilities, aiming at mitigating, detecting, or preventing the attacks. Fig. 11 illustrates the relation between vulnerabilities, attacks and protection mechanisms.

For more explanation, Table VI (a and b) elaborates the details of the major security requirements (i.e. CIA) in relation to vulnerabilities, attacks, and protection mechanisms.

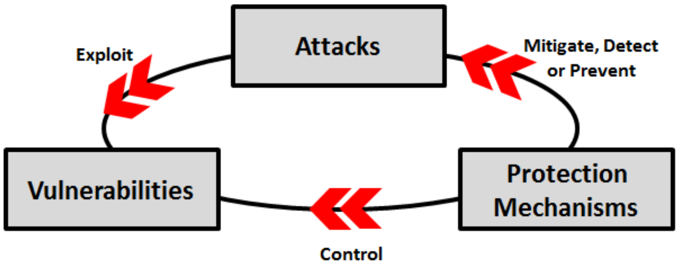


Fig. 11. Relation between Vulnerabilities, Attacks and Protection Mechanisms.

TABLE. VI. (A) CIA IN RELATION TO VULNERABILITIES, ATTACKS, AND PROTECTION MECHANISMS

CIA / Term	1 Confidentiality	2 Integrity		3 Availability	
Vulnerabilities	$v_{1,1}$	$v_{1,2}$		$v_{1,3}$	$v_{1,4}$
Attacks	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	
Protection Mechanisms	$pm_{1,1}$	$pm_{1,2}$	$pm_{1,3}$	$pm_{1,4}$	

(B) DESCRIPTION OF CODES

Code	Description	Code	Description	Code	Description
$v_{1,1}$	Disclosure to unauthorized party.	$a_{1,1}$	Eavesdropping attack	$pm_{1,1}$	Encryption
$v_{1,2}$	Modification by unauthorized party.	$a_{1,2}$	Tampering attack	$pm_{1,2}$	Access control
		$a_{1,3}$	Man-in-the-middle attack	$pm_{1,3}$	Authentication
$v_{1,3}$	Authorized parties are not able to access data.	$a_{1,4}$	DoS attack	$pm_{1,4}$	Hashing and digital signature
$v_{1,4}$	Natural disasters				

C. Classification of Attacks

In the mobile agent-based systems, attacks can be classified based on the nature of the attacks or based on the victim of the attack, as shown in Fig. 12.

According to the nature of the attack, attacks can be passive or active. In passive attacks, the attacker collects some information about the victim to be used later for malicious purposes. Therefore, no updates can affect the resources of the system. In active attacks, the attacker modifies the resources of the system so that there will be direct damage. As a result, active attacks are more dangerous than passive attacks. According to the victim of the attack, malicious agents can attack the operation execution on the platform, and in contrast, malicious platforms can attack the visiting mobile agents. Attacking the visiting mobile agents by platforms is more dangerous than attacking the platform by agents, which is because the platform has full control of the execution of the visiting agents within its operational environment. Table VII summarizes the most common possible attacks that agent-based systems suffer from, as well as their types.

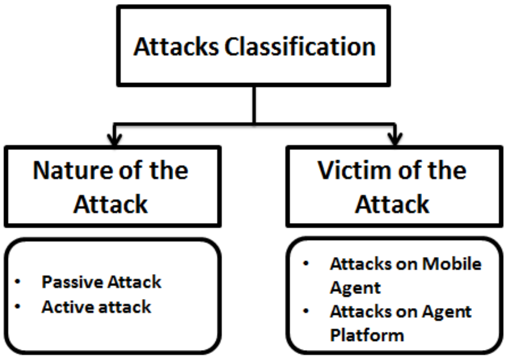


Fig. 12. Classification of Attacks in Agent-based Systems.

TABLE. VII. POSSIBLE ATTACKS

Victim of attack	Possible attacks	Nature of attack	
		Active	*Passive*
Mobile agent	1- DoS attack by the host of the agent	√	
	3- Eavesdropping on an agent's activities		√
	3- Blocking attack by the host	√	
	4- Modification of an agent by the host	√	
	5- Multiple colluded attack by hosts	√	
Agent platform	1- DoS attack with overmuch requests or exhausting the platform's memory or resources	√	
	2- Unauthorized access attack for: * shutting down platform * modifying policy file * performing any malicious activity	√	

D. Overview of Attacks

For the attacks that target the agent platform by a malicious agent, Fig. 13 and Fig. 14 illustrate the mechanisms by which the DoS and unauthorized access attacks are performed.

As shown in Fig. 13, a malicious agent migrates from an HM to a DM, asking to execute infinite requests of the same mission, so that the DM goes through an infinite loop of execution, thereby resulting in exhausting the resources of the host machine. In this case, if any other agent that exists in the DM asks for execution of its own mission, it is forced to wait forever, due to the allocation of the DM's resources for the malicious agent [79].

In Fig. 14, a malicious agent gains unauthorized access to a DM by exploiting some gaps in the system. After accessing the DM, the malware carried by the agent is then executed to damage some critical system files [80].

For the attacks that target the agent itself by a malicious DM, the DoS attack has a special case, as shown in Fig. 15.

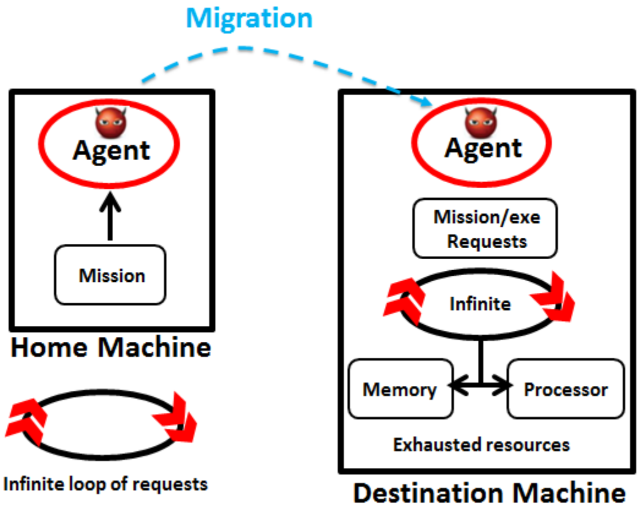


Fig. 13. Concept of the DoS Attack.

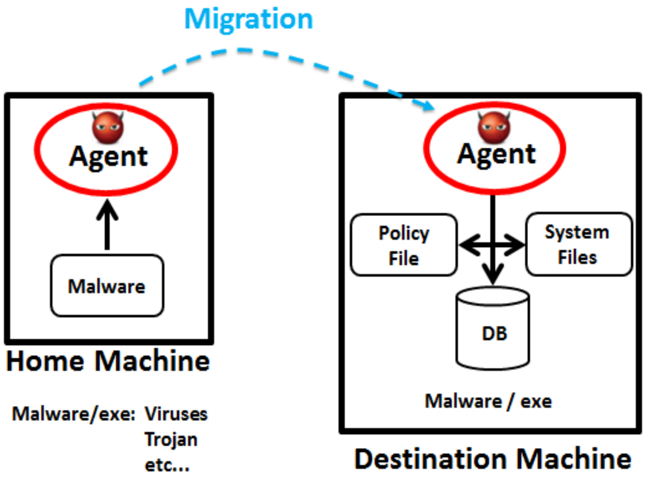


Fig. 14. Concept of the unauthorized Access Attack.

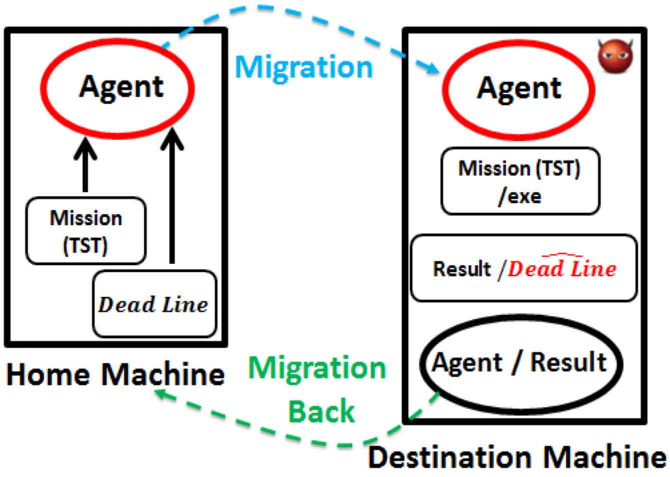


Fig. 15. Special Case of the DoS Attack.

In the DoS attack, the mobile agent carries a mission (time sensitive task TST), for which its execution time is restricted by a deadline ($T_{\text{Dead Line}}$). After the mobile agent migrates to the DM, the TST is executed there. The malicious DM deliberately delays (or lengthens) the execution time of the TST, so that it exceeds the predefined $T_{\text{Dead Line}}$. It is worth mentioning that even when the malicious DM does not modify the result of the TST, the result will be invalid and unused when it is received by the HM [81, 82].

Eavesdropping on an agent is applied by the malicious DM by monitoring and recording the activities of the agent during the time in which the agent executes its mission [83]. In the worst case, the eavesdropping attack can be turned into a blocking attack, where the DM completely prevents the agent from execution after a short period of monitoring [84].

Fig. 16 shows how the modification attack is applied on the visiting mobile agent by the DM. The DM does not monitor or block the agent, such that it is allowed to execute smoothly, but after generating the results of the execution, the DM tampers or changes the results. In other words, the mobile agent migrates back to the HM carrying wrong results [85].

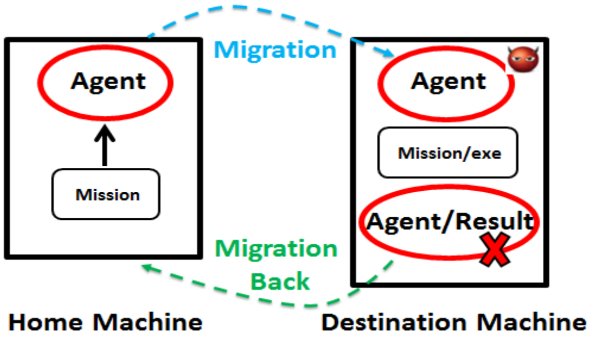


Fig. 16. Concept of the Modification Attack.

The modification attack becomes more dangerous when two or more DMs collude together to modify the results of the execution. This kind of attack is called the multiple colluded attack, as illustrated in Fig. 17.

In the multiple colluded attack, a series of (n) DMs $(DM_1, DM_2, \dots DM_{n-1}, DM_n)$ collude each other for a common malicious purpose, which is tricking the HM. Tricking the HM is achieved in such a way that: (1) two DMs, or more, modify the result generated after execution of the mission of the visiting mobile agent; and (2) all malicious DMs that modify the result use the same modification process [50]. Upon enacting this, the success of the multiple colluded attack can be adjusted by the following two conditions:

1) The original result $(\text{result}\,_i)$, generated at $(\text{DM}_i|0 < i \le n)$, is modified to be $(\widehat{\text{result}}_1)$.

2) At the HM, all the received results are modified, so that: $\widehat{\text{result}}_1 = \widehat{\text{result}}_2 = \cdots = \widehat{\text{result}}_{n-1} = \widehat{\text{result}}_n$

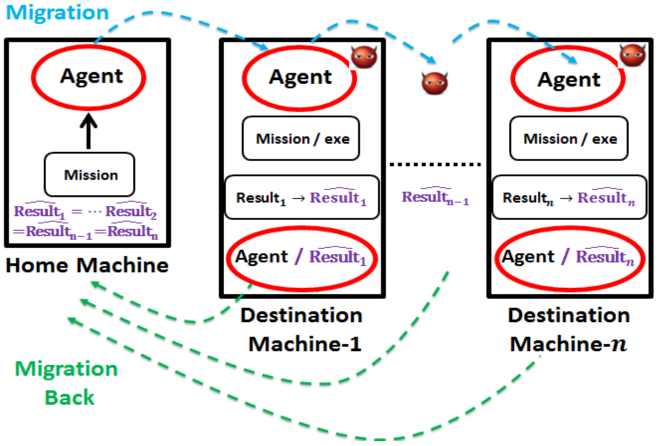

Migration
Agent
Mission
$\widehat{Result}_1 = \cdots \widehat{Result}_2 = \widehat{Result}_{n-1} = \widehat{Result}_n$
Home Machine
Agent
Mission / exe
$Result_1 \rightarrow \widehat{Result}_1$
Agent / $\widehat{Result}_1$
Destination Machine-1
$\widehat{Result}_{n-1}$
Agent
Mission / exe
$Result_n \rightarrow \widehat{Result}_n$
Agent / $\widehat{Result}_n$
Destination Machine-n
Migration
Back

Fig. 17. Concept of the Multiple Colluded Attack.

E. Maturity Model

To show the negative impact of the attacks, we propose a statistical model called the maturity model. The maturity model deals with the protection goals (i.e., code, status, and itinerary) as affected aspects of attacks. Since the protection goals are limited to the class of protecting the mobile agent compared to unlimited protection goals in the class of protecting the agent platform (i.e., any part of the system in the DM's platform can be a victim of attack), we deal only with those attacks that target the mobile agent as a victim (Table VII). Among the attacks that target the mobile agent as a victim (in Table VII), we consider only the DoS, modification, and multiple colluded attacks since they are considered as advanced attacks. Upon this consideration, in the maturity model, the DoS, modification, and multiple colluded attacks are considered as main criteria factors, while the protection goals are considered as affected aspects. All approaches contained in Table V above are evaluated. Our evaluation relies on three options to measure the negative impact of the criteria factors. Table VIII provides a description of the three used options.

TABLE. VIII. OPTIONS OF MEASUREMENT

Option	Description
√	When the factor has high negative impact.
×	When the factor has a low negative impact.
P	When the factor has a partially negative impact.

1) Analysis and discussion: Table IX below can be read horizontally or vertically, as illustrated in Fig. 18.

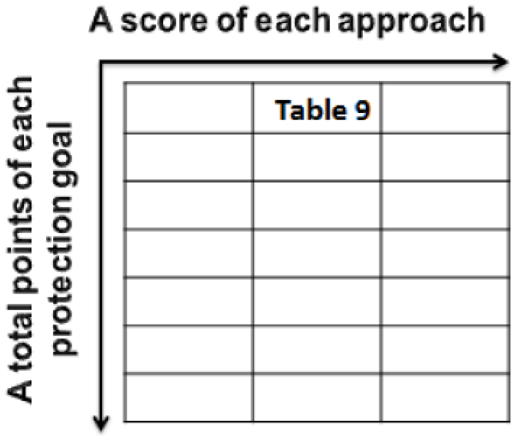


Fig. 18. Horizontal and Vertical Reading of Table IX.

TABLE. IX. EFFECT OF THE DOS ATTACK

Class	Term Technique	Protection Goals			Sub Totals		
		Code & Data	***State***	***Itinerary***	√	×	***P***
Protecting mobile Agent	***Collaborative agents***		Maturity: 2				
	[56]	P	√	×	1	1	1
	[57]	×	√	×	1	2	0
	RPE		Maturity: 2				
	[58]	P	√	×	1	1	1
	[59]	P	√	×	1	1	1
	Obfuscated code		Maturity: 1				
	[60]	P	√	×	1	1	1
	[61]	P	P	×	0	1	2
	Environment Key Generation		Maturity: 2				
	[62]	×	√	×	1	2	0
	[63]	×	√	×	1	2	0
	Execution tracking		Maturity: 2				
	[64]	P	√	×	1	1	1
	[65]	P	√	×	1	1	1
	Watermarking		Maturity: 3				
	[66]	√	√	×	2	1	0
	[67]	√	P	×	1	1	1
	Co-signing		Maturity: 0				
	[68]	P	P	×	0	1	2
	[69]	P	P	×	0	1	2
	Separation of Privileges		Maturity: 2				
	[70]	P	√	×	1	1	1
	[71]	P	√	×	1	1	1
	Fragmentation-based encryption		Maturity: 0				
	[72]	P	P	×	0	1	2
	[73]	P	P	×	0	1	2
Sub Totals:							
√	High negative impact	2	12	0	14		
×	Low negative impact	3	0	18	21		
P	Partial negative impact	13	6	0	19		
Total:					54		

If Table IX is read horizontally, then the numbers on the table represent the total points that each approach has obtained from all of the protection goals for each one of the three options. Each option has a score that varies in the range of [0, 1, 2, 3]. For instance, the corresponding numbers of the collaborative agents' technique are 2, 3, and 1 for the √, ×, and P options respectively. This in turn means that the DoS attack has a moderate impact on the approaches proposed under this technique because the score of the √ option equals 2. Thus, the maturity of the collaborative agents' technique is moderate under the threat of the DoS attack. A reasonable justification is that this technique was originally designed to protect the code itself, where assistant agents contribute to prevent the illegal redundancy of the same request (or the mission under execution). RPE, environment key generation, execution tracking, and separate of privileges techniques have the same level of maturity as the collaborative agents' technique. The maturity of the watermarking-based technique under the threat of the DoS attack is low because the score of the √ option equals 3, which is because the objective of this technique is to detect any modification in the code, not to prevent redundancy. The maturity of the obfuscated code-based technique under the threat of the DoS attack is high because the score of the √ option equals 1. The score is 1 because the obfuscation of the code prevents the attacker from extracting the real code so that it can be redundant. For the co-signing and fragmentation-based encryption techniques, the maturity level under the threat of the DoS attack is very high since the score of the √ option in each one equals 0. The score is 0 because the code is highly protected against being decrypted, and then against being exploited for redundancy. Within four groups, Table X ranks the previous techniques according to their maturity levels.

If Table IX is read vertically, then the numbers represent the total points that each protection goal has obtained for each one of the three options and is related to all of the approaches provided in protecting the mobile agent class. From the numbers that appear in Table IX, it can be noticed that the total number of points that the (√) option achieved is 14 points. These points distribute over the (code & data, state, and itinerary) protection goals with (2, 12, and 0) values respectively. For the (×) option, the protection goals achieved (3, 0, and 18) values from the sub-total points, with a total of 21. The corresponding values related to the protection goals for the (P) option are (13, 6, and 0) from the sub-total points, with a total of 19. In terms of percentages, Fig. 19(a) above shows the negative impact of the DoS attack's threat on the protection goals of all techniques (i.e., overall maturity against the DoS attack).

In Fig. 19(a), the high negative impact option (√) has the lowest percentage (0.25), while the low negative impact option (×) has the highest percentage. This in turn reflects a good maturity of the security approaches against the DoS attack. The reason behind this is that most of the approaches in all techniques are designed to protect the code of the agent.

Regarding modification and multiple colluded attacks, we rebuilt Table IX, scanned it vertically, and determined the percentages, as shown in Fig. 19(b), (c) above, respectively.

According to Fig. 19(b) and (c), most of the approaches suffer from the modification attack, with the percentage equal to 0.73 for the high negative impact option, which indicates a low overall maturity level. Compared to the modification attack, the security approaches have very poor maturity against the multiple colluded attack, with the percentage equal to 0.92 for the high negative impact option. The reason behind this is that both the state (which contains the results of the mission execution) and the code can be modified by a series of malicious DMs during the itinerary of the mobile agent.

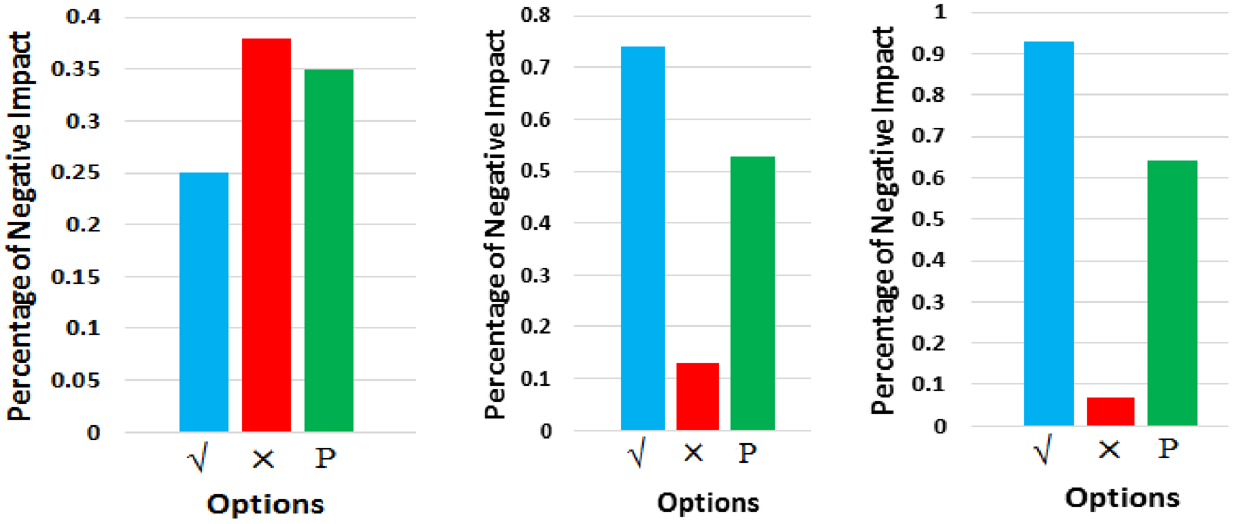
Based on the analysis derived from Fig. 19, Table XI ranks the attacks according to their danger on the visiting mobile agents.

TABLE. X. RANKING TECHNIQUES ACCORDING TO MATURITY LEVEL

Group	Techniques	Score of √ option	Maturity level
G1	Co-signing, fragmentation-based encryption	0	Very high
G2	Obfuscated code	1	High
G3	Collaborative agents, RPE, environment key generation, execution tracking, separation of privileges	2	Moderate
G4	Watermarking	3	Low

TABLE. XI. DANGER-BASED RANKING ATTACKS

Attack	Percentage of high negative impact option (√)	Ranking
Multiple colluded attack	0.92	First
Modification attack	0.73	Second
DoS attack	0.25	Third



(a) Overall Negative Impact of DoS attack (b) Maturity Against Modification attack (c) Maturity Against Multiple Colluded attack.

Fig. 19. Overall Negative Impact of Modification and Multiple Colluded Attacks.

VII. Evaluation of Security of Agent-Based Systems

Designing a secure agent-based system requires defining a threat model at the beginning. The clear threat model contains four parts, as shown in Fig. 20.

The first part of the threat model is the attacker. In agent-based systems, the attacker can be the mobile agent, the destination machine, or any malicious party in the system such as the man-in-the-middle. After deciding who the attacker is, it determines his objective, which is the second part on the threat model. The objective of the attacker can be attacking one or more of the protection goals if the victim is the mobile agent or applying any malicious activity if the victim is the agent platform. Based on the determined objective, the type of the attack is defined to be active or passive. Finally, the capability of the attacker is listed in the context of the attacks that the attacker can apply, such as eavesdropping, modification, DoS, or multiple colluded attacks, etc.

After defining the threat model, the agent-based system is built so that defenses (or security controls) against the capability of the attacker are implemented. After building the system, it undergoes a validation process to ensure that the security controls are able to detect or prevent the attacks. Finally, to measure the efficiency of the security controls, security metrics-based evaluation must be performed. In this context, we explore the security metrics used in agent-based systems. We classify the security metrics into three main categories, as shown in Fig. 21.

For the first group, the number of security requirements that are achieved is used as a metric to evaluate the proposed agent-based system. Therefore, the strength of the system is linked to satisfying a higher number of security requirements. An available tool, called Scyther [86], can be employed in this group to check if a certain security requirement is satisfied or not.

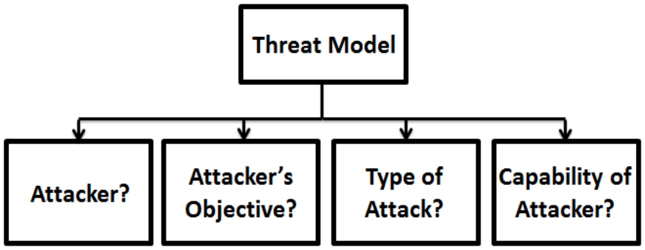


Fig. 20. Parts of the Threat Model.

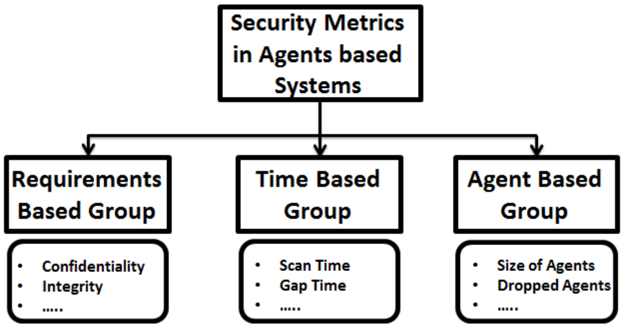


Fig. 21. Classification of Security Metrics.

For the second group, time is mainly employed in different forms as a metric. Such forms are: (1) the time needed for scanning the code of the agent; (2) the time needed for encryption and decryption; (3) the time needed for calculation of the hash of the agent's code; and (4) the time gap, which is the time consumed between the attack detection and the actual update that is applied to the system.

For the third group, the size of the agent can be used as a metric as well as the events that occur in the system and are related to the agent, such as the number of agents that are dropped by the platform, and the number of agents that are attacked in time units.

VIII. Challenges and Research Questions

In this section, we provide the challenges encountered by the corresponding research questions (five questions) according to the section presented in this work. In addition, we present an additional two research questions as inferred ones.

1) According to Section III (Cyber Security requirements in the lifecycle of mobile agent), requirements related to security and those related to privacy are the challenges. The two kinds of requirements should be achieved in agent-based systems to state that we have a comprehensive secure system in terms of cyber security. However, researchers distinguish between security and privacy and rank security at the top compared to privacy. Therefore, how to satisfy all the security requirements simultaneously in an agent-based system is the first research question.

2) According to Section IV (Classification of security techniques in mobile agents), the visiting mobile agent my damage the destination machine, and at the same time, the visited destination machine may attack the incoming agent. This leads to the challenge that can be represented by the second research question: how to ensure that the visiting mobile agent and the host machine do not attack each other so that each party operates in a completely secure manner.

3) According to section VI (Protection goals and attacks), two research questions are motivated as follows:

- Since one of the most important benefits of mobile agents is reducing the network overhead as well as solving the transmission challenge problem, protecting the state (which carries the results of execution back to the home machine) is a major challenge. Therefore, the research question related to how to guarantee the integrity of the results is a critical one to be answered.

- Regarding attacks, the major concern occurs when the destination machine tries to apply advanced attacks on the visiting mobile agent, such as modification, DoS, and multiple colluded attacks. Therefore, how to build strong protection mechanisms that ensure high resistance against such attacks is the corresponding research question.

4) According to Section VII (evaluation of security of agent-based systems), ensuring efficient and standard security metrics related to the protection mechanisms (i.e., out of the three groups mentioned in Fig. 21) is a challenge. Therefore, how to quantify the security of agent-based systems by an efficient mathematical model is an important aspect for the evaluation process. This will be a strong point with respect to comparing different protection mechanisms.

5) It is argued that the need of an effective protection mechanism is mandatory. However, how to equip the agent with a protection mechanism that provides a self-protection feature is another research question.

6) Since the destination machine has full control over the visiting mobile agent, it is important from the security point of view to isolate communications with the destination machine. In other words, during the performance of the agent's mission within the operational environment of the destination machine, how to endow the agent with a self-communication feature is an overarching research question.

IX. Conclusion

Compared to other software technologies, agent-based software technology presents itself as an effective solution for many problems in distributed systems, such as network overhead and transmission challenge. However, the security issue is a main factor that contributes to limitations of the benefits of agent-based software technology as well as its applications. The main reason behind this issue is that the agents can be attacked by the destination machines where they perform the missions, or the visiting agents can perform malicious activities on the host machine. Moreover, advanced attacks such as DoS, modification, and multiple colluded attacks can exacerbate the security problem. Based on the attacker (the visiting mobile agent and the destination or host machine), we review different techniques used to ensure the security in agent-based systems, critique them, and compare them according to well-defined cyber security requirements (in both the security and privacy aspects). Based on protection goals (code and data, state, and itinerary of the mobile agent), a maturity model is employed to analyse the security techniques as well as rank the strength of the attacks. Finally, seven research questions are provided in the research field of agent security that should be answered to ensure comprehensive security in agent-based systems.

References

[1] Padgham, Lin, and John Thangarajah. "Agent Oriented Software Engineering: Why and How." VNU Journal of Science: Natural Sciences and Technology 27.3 (2016).

[2] Qiu, Linrun, and Kangshun Li. "The Research of Intelligent Agent System Architecture Based on Cloud Computing." Computational Intelligence and Security (CIS), 2016 12th International Conference on. IEEE, 2016.

[3] [Caglayan, Alper, and Colin Harrison. "Agent sourcebook." (2011).

[4] Satoh, Ichiro. "Building reusable mobile agents for network management." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 33.3 (2003): 350-357.

[5] Bieszczad, Andrzej, Bernard Pagurek, and Tony White. "Mobile agents for network management." IEEE Communications Surveys 1.1 (1998): 2-9.

[6] Bergenti, Federico, Eleonora Iotti, and Agostino Poggi. "Core features of an agent-oriented domain-specific language for JADE agents." Trends in Practical Applications of Scalable Multi-Agent Systems, the PAAMS Collection. Springer, Cham, 2016. 213-224.

[7] Urra, Oscar, et al. "Mobile agents and mobile devices: Friendship or difficult relationship?." (2009). J. Phys. Agents 3, 2 (2009), 27−37.

[8] Mobil Agent Computing website, online, available: https://www.cis.upenn.edu/~bcpierce/courses/629/papers/Concordia-WhitePaper.html , (2018), 6th October.

[9] Jade website, online, available: http://jade.tilab.com/ , (2018), 6th October.

[10] Binu A website, online, available: http://csr.cusat.ac.in/people/binua /blog/ibm-aglets-workbench-installation-agent-programming,(2018), 6th October.

[11] Wu, Bing, et al. "A survey of attacks and countermeasures in mobile ad hoc networks." Wireless network security. Springer, Boston, MA, 2007. 103-135.

[12] Karnik, Neeran M., and Anand R. Tripathi. "A security architecture for mobile agents in Ajanta." Distributed Computing Systems, 2000. Proceedings. 20th International Conference on. IEEE, 2000.

[13] Bakre, Ajay, and B. R. Badrinath. "M-RPC: A remote procedure call service for mobile clients." Proceedings of the 1st annual international conference on Mobile computing and networking. ACM, 1995.

[14] Stamos, James W., and David K. Gifford. "Implementing remote evaluation." IEEE Transactions on Software Engineering 16.7 (1990): 710-722.

[15] Carzaniga, Antonio, Gian Pietro Picco, and Giovanni Vigna. "Designing distributed applications with mobile code paradigms." Proceedings of the 19th international conference on Software engineering. ACM, 1997.

[16] Ali, Anwaar, et al. "Big data for development: applications and techniques." Big Data Analytics 1.1 (2016): 2.

[17] Ramírez-Gallego, Sergio, et al. "Big data: tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce." Information Fusion 42 (2018): 51-61.

[18] Boubiche, Sabrina, et al. "Big Data Challenges and Data Aggregation Strategies in Wireless Sensor Networks." IEEE Access 6 (2018): 20558-20571.

[19] Zhu, Li, et al. "Big Data Analytics in Intelligent Transportation Systems: A Survey." IEEE Transactions on Intelligent Transportation Systems (2018).

[20] Kavak, Hamdi, et al. "Big data, agents, and machine learning: towards a data-driven agent-based modeling approach." Proceedings of the Annual Simulation Symposium. Society for Computer Simulation International, 2018.

[21] Karami, Mahtab, and Ali HOSSEINI SHAHMIRZADI. "Applying Agent-based Technologies in Complex Healthcare Environment." Iranian journal of public health 47.3 (2018): 458.

[22] Kaeri, Yuki, et al. "Agent-Based System Architecture Supporting Remote Collaboration via an Internet of Multimedia Things Approach." IEEE Access 6 (2018): 17067-17079.

[23] Sun, Weiwei, et al. "An air index for spatial query processing in road networks." IEEE Transactions on Knowledge and Data Engineering 27.2 (2015): 382-395.

[24] Alrahhal, Mohamad Shady, Maher Khemekhem, and Kamal Jambi. "Achieving load balancing between privacy protection level and power consumption in location based services." (2018).

[25] Lange, Danny B., and Mitsuru Oshima. "Seven good reasons for mobile agents." Communications of the ACM 42.3 (1999): 88-89.

[26] Chaabouni, Taha, and Maher Khemakhem. "Resource Management Based on Agent Technology in Cloud Computing." Advances in Information Technology for the Holy Quran and Its Sciences (32519), 2013 Taibah University International Conference on. IEEE, 2013.

[27] Arfat, Yasir, and Fathy Elbouraey Eassa. "A Survey on Fault Tolerant Multi Agent System." IJ Inf. Technol. Comput. Sci 9 (2016): 39-48.

[28] Madkour, Mohamed A., et al. "Mobile Agent Framework for Distributed Network Performance Management." International Journal of Computer Applications 88.5 (2014).

[29] Imran, Muhammad, Fathy Eassa, and Kamal Jambi. "Using Agent Technology for Security Testing of WEB Based Applications." SEDE–2015: 3-10.

[30] Alrahhal, Mohamad Shady, Maher Khemakhem, and Kamal Jambi. "Agent-Based System for Efficient kNN Query Processing with Comprehensive Privacy Protection." International Journal Of Advanced Computer Science And Applications 9.1 (2018): 52-66.

[31] Wernke, Marius, et al. "A classification of location privacy attacks and approaches." Personal and ubiquitous computing18.1 (2014): 163-175.

[32] Alrahhal, Mohamad Shady, Maher Khemakhem, and Kamal Jambi. "A Survey on Privacy of Location-Based Services: Classification, Inference Attacks, And Challenges." Journal of Theoretical & Applied Information Technology 95.24 (2017).

[33] Alrahhal, Mohamad Shady, et al. "AES-Route Server Model for Location based Services in Road Networks." International Journal Of Advanced Computer Science And Applications 8.8 (2017): 361-368.

[34] Kharaji, Morteza Yousefi, and Fatemeh Salehi Rizi. "A fast survey focused on methods for classifying anonymity requirements." International Journal of Computer Science and Information Security 12.4 (2014): 59.

[35] Deng, Mina. "Privacy Preserving Content Protection (Privacy behoud content protection)." (2010).

[36] Wahbe, Robert, et al. "Efficient software-based fault isolation." ACM SIGOPS Operating Systems Review. Vol. 27. No. 5. ACM, 1994.

[37] Van't Noordende, Guido, Frances MT Brazier, and Andrew S. Tanenbaum. "A security framework for a mobile agent system." Proceedings of the 2nd International Workshop on Security in Mobile Multiagent Systems (SEMAS 2002), associated with AAMAS-2002, Bologna, Italy. 2002.

[38] Rights, Retains Full. "Secure Coding. Practical steps to defend your web apps." (2007).

[39] Malik, Najmus Saqib, and Albert Treytl. "Optimizing Security Computation Cost for Mobile Agent Platforms." Industrial Informatics, 2007 5th IEEE International Conference on. Vol. 1. IEEE, 2007.

[40] Cooper, David, et al. Security Considerations for Code Signing. No. OTHER-. 2018.

[41] Appel, Andrew W., and David McAllester. "An indexed model of recursive types for foundational proof-carrying code." ACM Transactions on Programming Languages and Systems (TOPLAS) 23.5 (2001): 657-683.

[42] Sekar, R., et al. "Model-carrying code: a practical approach for safe execution of untrusted applications." ACM SIGOPS Operating Systems Review. Vol. 37. No. 5. ACM, 2003.

[43] Kim, Hyong-Soon, and Eunyoung Lee. "Verifying Code toward Trustworthy Software." Journal of Information Processing Systems 14.2 (2018).

[44] Borselius, Niklas. "Mobile agent security." Electronics & Communication Engineering Journal 14.5 (2002): 211-218.

[45] Chess, David, et al. "Itinerant agents for mobile computing." IEEE Personal Communications 2.5 (1995): 34-49.

[46] Ordille, Joann J. "When agents roam, who can you trust?." Emerging Technologies and Applications in Communications, 1996. Proceedings., First Annual Conference on. IEEE, 1996.

[47] Guan, Sheng-Uei, Tianhan Wang, and Sim-Heng Ong. "Migration control for mobile agents based on passport and visa." Future Generation Computer Systems 19.2 (2003): 173-186.

[48] Marikkannu, P., R. Murugesan, and T. Purusothaman. "AFDB security protocol against colluded truncation attack in free roaming mobile agent environment." Recent Trends in Information Technology (ICRTIT), 2011 International Conference on. IEEE, 2011.

[49] Hefeeda, Mohamed, and Bharat Bhargava. "On mobile code security." Purdue University (Oct. 2001) (2001).

[50] Venkatesan, S., C. Chellappan, and P. Dhavachelvan. "Performance analysis of mobile agent failure recovery in e-service applications." Computer Standards & Interfaces 32.1-2 (2010): 38-43.

[51] Venkatesan, S., et al. "Artificial immune system based mobile agent platform protection." Computer Standards & Interfaces35.4 (2013): 365-373.

[52] Tsiligiridis, Theodore A. "Security for mobile agents: Privileges and state appraisal mechanism." Neural Parallel And Scientific Computatiions 12.2 (2004): 153-162.

[53] Farmer, William M., Joshua D. Guttman, and Vipin Swarup. "Security for mobile agents: Authentication and state appraisal." European Symposium on Research in Computer Security. Springer, Berlin, Heidelberg, 1996.

[54] Bagga, Pallavi, Rahul Hans, and Vipul Sharma. "N-grams Based Supervised Machine Learning Model for Mobile Agent Platform Protection against Unknown Malicious Mobile Agents." International Journal of Interactive Multimedia & Artificial Intelligence 4.6 (2017).

[55] Bagga, Pallavi, Rahul Hans, and Vipul Sharma. "A Biological Immune System (BIS) inspired Mobile Agent Platform (MAP) security architecture." Expert Systems with Applications 72 (2017): 269-282.

[56] Roth, Volker. "Mutual protection of co-operating agents." Secure Internet Programming. Springer, Berlin, Heidelberg, 1999. 275-285.

[57] Madkour, A. M., et al. "Securing mobile-agent-based systems against malicious hosts." World Applied Sciences Journal 29.2 (2014): 287-297.

[58] Yee, Bennet S. "A sanctuary for mobile agents." Secure Internet Programming. Springer, Berlin, Heidelberg, 1999. 261-273.

[59] Young, Adam, and Moti Yung. "Sliding encryption: a cryptographic tool for mobile agents." International Workshop on Fast Software Encryption. Springer, Berlin, Heidelberg, 1997.

[60] Hohl, Fritz. "Time limited blackbox security: Protecting mobile agents from malicious hosts." Mobile agents and security. Springer, Berlin, Heidelberg, 1998. 92-113.

[61] Badger, Lee, et al. "Self-protecting mobile agents obfuscation techniques evaluation report." Network Associates Laboratories, Report (2002): 01-036.

[62] Riordan, James, and Bruce Schneier. "Environmental key generation towards clueless agents." Mobile agents and security. Springer, Berlin, Heidelberg, 1998. 15-24.

[63] Tschudin, Christian. "Apoptosis—The programmed death of distributed services." Secure Internet Programming. Springer, Berlin, Heidelberg, 1999. 253-260.

[64] Tan, Hock Kim, and Luc Moreau. "Extending execution tracing for mobile code security." Proceedings of Second International Workshop on Security of Mobile MultiAgent Systems (SEMAS'2002). 2002.

[65] Tan, Hock Kim, and Luc Moreau. "Certificates for mobile code security." Proceedings of the 2002 ACM symposium on Applied computing. ACM, 2002.

[66] Esparza, Oscar, et al. "Mobile agent watermarking and fingerprinting: tracing malicious hosts." International Conference on Database and Expert Systems Applications. Springer, Berlin, Heidelberg, 2003.

[67] Esparza, Oscar, et al. "Punishing manipulation attacks in mobile agent systems." Global Telecommunications Conference, 2004. GLOBECOM'04. IEEE. Vol. 4. IEEE, 2004.

[68] Cheng, Jeff SL, and Victor K. Wei. "Defenses against the truncation of computation results of free-roaming agents." International Conference on Information and Communications Security. Springer, Berlin, Heidelberg, 2002.

[69] Linna, Fan, and Liu Jun. "A free-roaming mobile agent security protocol against colluded truncation attack without trusted third party." Business Management and Electronic Information (BMEI), 2011 International Conference on. Vol. 2. IEEE, 2011.

[70] Al-Jaljouli, Raja, and Jemal H. Abawajy. "Secure mobile agent-based e-negotiation for on-line trading." Signal Processing and Information Technology, 2007 IEEE International Symposium on. IEEE, 2007.

[71] Traut, Eric, et al. "Protection agents and privilege modes." U.S. Patent No. 8,380,987. 19 Feb. 2013.

[72] Srivastava, Shashank, and G. C. Nandi. "Fragmentation based encryption approach for self protected mobile agent." Journal of King Saud University-Computer and Information Sciences 26.1 (2014): 131-142.

[73] El Rhazi, Abdelmorhit, Samuel Pierre, and Hanifa Boucheneb. "A secure protocol based on a sedentary agent for mobile agent environments." Journal of Computer Science 3.1 (2007): 35-42.

[74] Arepalli, Abhishek, S. Srinivasa Rao, and P. Jagadeeshwara Rao. "A Spatial Disaster Management Framework for Smart Cities—A Case." Proceedings of International Conference on Remote Sensing for Disaster Management: Issues and Challenges in Disaster Management. Springer, 2018.

[75] Hartama, D., et al. "Smart City: Utilization of IT resources to encounter natural disaster." Journal of Physics: Conference Series. Vol. 890. No. 1. IOP Publishing, 2017.

[76] Abid, A., A. Kachouri, and A. Mahfoudhi. "Data analysis and outlier detection in smart city." Smart, Monitored and Controlled Cities (SM2C), 2017 International Conference on. IEEE, 2017.

[77] Liu, Ce, et al. "Motion magnification." ACM transactions on graphics (TOG). Vol. 24. No. 3. ACM, 2005.

[78] Wadhwa, Neal, et al. "Phase-based video motion processing." ACM Transactions on Graphics (TOG) 32.4 (2013): 80.

[79] Mittal, Praveen, and Manas Kumar Mishra. "Trust and Reputation-Based Model to Prevent Denial-of-Service Attacks in Mobile Agent System." Towards Extensible and Adaptable Methods in Computing. Springer, Singapore, 2018. 297-307.

[80] Samet, Donies, Farah Barika Ktata, and Khaled Ghedira. "Securing Mobile Agents, Stationary Agents and Places in Mobile Agents Systems." KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications. Springer, Cham, 2018.

[81] Wood, Anthony D., and John A. Stankovic. "A taxonomy for denial-of-service attacks in wireless sensor networks." Handbook of sensor networks: compact wireless and wired sensing systems (2004): 739-763.

[82] Greenberg, Michael S., Jennifer C. Byington, and David G. Harper. "Mobile agents and security." IEEE Communications magazine 36.7 (1998): 76-85.

[83] Prem, M. Vigilson, and S. Swamynathan. "Securing mobile agent and its platform from passive attack of malicious mobile agents." IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-2012). IEEE, 2012.

[84] Singh, Rajwinder, and Mayank Dave. "Rescuing data of mobile agents blocked by malicious hosts in e-service applications." 2011 International Conference on Multimedia, Signal Processing and Communication Technologies. IEEE, 2011.

[85] Mitrovic, Nikola, and Unai Arronategui Arribalzaga. Mobile Agent security using Proxy-agents and Trusted domains. No. INPRO--2009-035. 2002.

[86] Cremers, Cas JF. "The scyther tool: Verification, falsification, and analysis of security protocols." International Conference on Computer Aided Verification. Springer, Berlin, Heidelberg, 2008.

Indoor Positioning System using Regression-based Fingerprint Method

Reginald Putra Ghozali[1], Gede Putra Kusuma[2]
Computer Science Department, BINUS Graduate Program–Master of Computer Science
Bina Nusantara University, Jakarta, Indonesia, 11480

***Abstract*—Indoor Positioning System has opportunity to be used in different business platform. Based on past research, optimized localization method for Bluetooth Low Energy (BLE) to predict position of person or object with high accuracy has not been found yet. Most recent research that have solve Received Signal Strength (RSS) inconsistent value is using fingerprint method. This paper proposed a deep regression machine learning using convolutional neural network (CNN) with regression-based fingerprint model to estimate real position. The model used 5 nearest fingerprints as reference RSS values with their location (x or y) label as inputs to produce output of single value position (x or y), then repeat the process to produce second value of position to create complete coordinate of estimated position. To evaluate the proposed model, a comparison between training data with validation data using Root Mean Squared Error (RMSE) is used. The comparisons are with Multilayer Perceptron model and with the weighted sum method as benchmark. The experiment Gave results of mean distance and 90th percentile distance between proposed model with the benchmark. CNN model achieved accuracies of lower than 330cm at 90th percentile with mean distance lower than 185cm. Weighted sum model achieved accuracies lower than 360cm at 90th percentile with mean distance higher than 185cm, and MLP is in between them. The result demonstrates that the proposed method outperformed the benchmark methods.**

***Keywords*—Indoor positioning system; fingerprinting; regression machine learning; convolutional neural network**

I. Introduction

Indoor Positioning System (IPS) has been a trend in research even until now. The system is capable of diverse purposes. Different from Global Positioning System (GPS) that optimized for outdoor environment where mostly no significant obstructions, IPS needs to work for indoor environment that have significant obstructions like walls, roof, and every room object including human itself. There are more demands of accurate position for IPS.

Various surveys already written for IPS related topics [1] [2] from outdated methods are still used to more recent methods and most of them did not differentiate the methods for Wi-Fi or Bluetooth Low Energy (BLE). While they use similar bandwidth (2.4GHz radio frequency), the actual systems have different computing capabilities and availability. Among them, BLE has been used frequently by reasons of low cost, very low battery consumption, and high availability as supported by most modern smartphones.

BLE used 2.4 GHz unlicensed frequency (2402 to 2480 MHz) with total of 40 channels (2 MHz for each channel width). And using 3 channels for discovery services (channel 37, 38, 39) [3] [4]. Many algorithms have been used for optimizing accuracy of the system. Such as multilateration and fingerprinting. Even so, there is not yet optimized solutions for high accuracy using BLE technology [1].

There are many factor that affect the BLE radio propagation of the signals in indoor environments as BLE using radio signals, e.g., multipath effect, causing a random behavior in the Received Signal Strength (RSS) measurements caused by reflection [5], movement rate of human [6], and fast fading when measuring within a little time [7]. To solve these problems, fingerprinting method is needed to estimate indoor position that needs estimation algorithm to ensure accuracy of position.

To get object’s location based on received signal strength from BLE, certain measurement method is needed. Current popular method is fingerprinting. Where localization algorithms used for measure or estimate location. It consists at least 2 steps: Offline step and Online step. Offline step used to create a radio mapping of possible location from given signal strength received. While online step [1] will match the received signals during online moments with radio mapping from previous step to determine object’s location. Method to determine estimated position will affect the accuracy of estimated real position of an object. Different methods have been used, started from K-Nearest Neighbor (KNN) to using machine learning.

Current state-of-the-art [8] is using Polynomial regression to calculate distance as propagation model. Where RSS received processed using weighted centroid localization or weighted sum to get coordinate and using polynomial regression model to get distance. Both are calculated using RSS signal from 3 advertisement channels of each beacon. Then both results filtered using outlier detection to clean the result, by combine fingerprinting with polynomial regression model distance into combined distance. This filtered result will be processed using extended Kalman filtering using filtered distance from first outlier detection. Result from extended Kalman filtering will be filtered again using outlier detection to remove false measurement. This result then processed again with extended Kalman filtering into estimated position that will be compared with radio map to get the real position. This method is using distance-based measurement. Where the error rate is pretty high, caused by multipath effect and person movement rate, which is why the distance is filtered through many processes mentioned in the method. Other weakness is it takes lots of calculation time.

This paper intended to implement probabilistic method of fingerprinting using Deep Learning Convolutional Neural Networks Regression Model to estimate position of a person. The proposed method is expected to improve accuracy of estimated position. The design consists of BLE beacons as signal transmitter, and mobile smartphone as signal receiver. Signal received in the device will be processed using fingerprinting method, then estimated by Convolutional Neural Networks (CNN) with self-designed architecture, resulting an estimated position of a person.

This paper is divided into 8 sections, starting with introduction. Section 2 provides overview of related works from past to current state-of-the-art. Section 3 describes detail of proposed method. Section 4 describes experimental design and data collection method. Section 5 presents the experimental results and comparison with other methods. Section 6 provides discussion on limitations and complexity issues. Section 7 summarizes our works. Suggestions of further research are provided in Section 8.

II. Related Works

Indoor positioning Systems have been made using different methods. A review paper [1] describes lots of technologies and techniques used in developing indoor positioning system. Most used technologies until now is radio frequency-based technology using BLE. On the techniques side, most used until now is fingerprinting. Fingerprinting consist of two phases: offline phase, where a method of “training” to create possible mapping of estimated position that called as radio map from RSS. And online phase, where real position of an object is estimated by matching the received RSS with radio map using localization algorithms. At this phase, different kind of methods are used to optimize the estimated position. Few examples that described in this paper are summarized in Table I.

TABLE I. Summary of Recent Studies on Indoor Positioning Fingerprinting

Authors (Year)	**Inputs Variable(s)**	**Output**	**Method(s)**	**Performance**	**Result(s)**
Yu, et al. (2014)	5 WiFi RSS	Estimated position (x,y)	Cluster K-nearest Neighbor Manhattan Distance	Localization accuracy Between 2.4G and 5G WiFi signal.	1.4700 for 2.4G 1.1500 for 5G
Li, et al. (2016)	8 WiFi RSS	Estimated position (x,y)	Weighted K-nearest Neighbor Improved Manhattan Distance	Cummulative Distributive Error at 80th percentile.	2.10m for Euclidean distance 1.88m for Manhattan distance 1.48m for improved Manhattan distance
Faragher, R., & Harle, R. (2014)	19 BLE RSS	Estimated position (x,y)	Gaussian Process Regression Bayesian Likelihood Function Maximum a posteriori probability Euclidean Distance	Cumulative Probability / Distributive Error Between WiFi and BLE	8.5m at 95th percentile of the time for WiFi 2.6m at 95th percentile of the time for BLE
Faragher, R., & Harle, R. (2015)	19 RSS from BLE and 3 RSS from WiFi	Estimated position (x,y)	Proximity algorithm Weighted KNN Gaussian Process Regression Euclidean Distance	Cumulative Probability / Distributive Error Between WiFi and BLE	<3m at 95th percentile of the time for BLE <6m at 95th percentile of the time for WiFi
Zhuang, Yang, Li, Qi, & El-Sheimy, (2016)	20 BLE RSS	Estimated position (x,y)	Weighted Centroid Localization Algorithm Polynomial Regression Model Propagation Model Outlier Detection Extended Kalmann Filtering	Cumulative Distributive Error Between Propagation Model and Regression Model	3.1m at 90th percentile for Regression Model 3.8m at 90th percentile for Propagation Model
Tuncer & Tuncer, (2015)	3 RSS and 3 ID from 4 BLE	Estimated position (x,y)	Artificial Neural Network (ANN) Centroid Localization Algorithm (CLA)	Hyperbolic Tangent Sigmoid and Linear Transfer Function for Cost Function. Root Mean Squared Error for performance.	Training: 22m for ANN 58.24m for CLA Testing: 33.26m for ANN 108.15m for CLA
Xu, Wu, Li, Zhu, & Wang, (2018)	49 RFID RSS	Estimated position (x,y)	Support Vector Regression-LANDMARC algorithm	Root Mean Squared Error	35.532m for LANDMARC 27.226m for SA-SVR-LANDMARC 26.936m for BP-LANDMARC 20.243m for SVR-LANDMARC

Localization algorithms in fingerprinting for indoor positioning system are used to determine a position of a person or object. Based on [9], localization algorithms can be divided into deterministic and probabilistic method. Deterministic methods use metric to measure signal and fingerprint location based on the data. Some advantages using these methods are easy to implement and usually low computation. However, as the accuracy can be improved using complex measurement and many access points, the computation can take longer. Most traditional method is K-nearest neighbor (KNN). Author in [10] using modified KNN called cluster-KNN with three nearest Manhattan distances from BLE signals estimate position. KNN use received signal strength indicator received from fingerprinting during offline phase of detecting signal to produce fingerprint map. These signals then processed with the algorithm using either Manhattan distance or Euclidean distance to classify nearest access point that can represent the person's or object's position who brought the emitter devices. While the algorithm itself is not too much complex, making computation far faster, it sacrificed accuracy of the positioning by taking access point location as the detected person's location. The research used 2.4G and 5G WiFi signal to compare localization accuracy from Manhattan distance of RSS average error, resulting 1.4700 for 2.4G and 1.1500 for 5G. Other KNN method used by [11] is called weighted-KNN. Where a parameter called weight assigned to every coordinate according to the value of distance. This paper used improved Manhattan distance, where certain constant used as threshold to consider increment of distance difference. This research used comparison between Manhattan distance and Euclidean distance using simulation software that simulate an office room with 8 rooms with an access point placed at innermost of the room and a corridor, resulting 1.88m positioning error at 80th percentile for Manhattan distance, and 2.10m positioning error at 80th percentile for Euclidean distance, and 1.48m at 80th percentile for improved Euclidean distance.

Probabilistic methods use estimation to determine position based on training set of signal data, and then choosing the most likely position of the target. Example of probabilistic methods is: Gaussian process [3] [4]. Gaussian Process in indoor positioning system used to estimate possibility from Bayes rule. Author in [3] used Gaussian process to point location based on uncertainty of Bayes rule estimation of received signal strength. Using 19 BLE beacons, they received accuracy of error rate around 2.6m in 95%, compared with WiFi, the accuracy of 8.5m in 95%. Author in [4] continues research of [3], with detailed and motivating reasons to use BLE for indoor positioning.

Author in [8] is using polynomial regression model to estimate cumulative distribution methods of average distance errors for each BLE beacon, then compare the result with same data using propagation model. The RSS data came from three advertisement channels processed through model using Fingerprinting for location and Polynomial Regression Model for distance resulting three different locations and three different distances. Then, each of them improved the distance estimation by using statistical method from first Outlier Detection. This improved distance estimation processed with Extended Kalman Filtering resulting estimated current target location. This result processed in second Outlier Detection to remove outliers and the outputs will be compared with RSS mapping database to select most appropriate location. Polynomial Regression Model used to calculate distance (1).

$$^{d}PRM = \sum_{i=0}^{n} c_i \,.\, RSS^{i} \tag{1}$$

Where c_i is coefficient from n-degree polynomial, then multiplied by RSS value. They use 20 beacons with 3 advertisement channels, resulting total of 60 average distance errors. The polynomial degree is 5. The result is, polynomial degree 2 through 5 have similar result and better than first polynomial and propagation model. With polynomial degree 2 has fastest computation than other. This paper stated that at 90% estimated error of used data using polynomial regression model is 3.1m, while using propagation model is 3.8m. Based from this paper, assumed that polynomial regression at degree 2 has high accuracy and fast computation.

Machine learning can be used on either classification problem or regression problem. One kind of machine learning type is Artificial Neural Network (ANN). ANN work similarly like human brain, just like neuron interconnected each other inside brain. The neuron part gives brain capability to learning, prediction, and recognition. This means ANN can be trained to learn something. Author in [12] used ANN for localization compared with centroid localization algorithm. Location of the user is estimated by using coordinates of at least three anchor points to calculate the central point, then using the distance between central point and location of user to find location error. They proposed three layers ANN model (input, hidden, output) using hyperbolic tangent sigmoid and linear transfer functions, with backpropagation algorithm for network training. Results of the RMSE from training are 22m for ANN and 58.24m for centroid localization algorithm. For testing results, 33.26m for ANN and 108.15m for centroid localization algorithm. Another kind of regression is Support Vector Regression (SVR) that uses supervised learning model to analyze a regression line model that represent all data closest to the plane. Author in [13] used SVR to improve RFID-based indoor positioning system. Using vector of RSS values read from single tag and reference position to train model using linear regression. The research compared RMSE results from different kind of LANDMARC algorithm, which is reference-tag based positioning system using RFID. This method consists of reference label matrix for positioning label in space, RSSI values from unknown position and reference labels, and KNN algorithm for positioning. The research resulted 25.532m for non-customized LANDMARC, 27.226m for SA-SVR-LANDMARC, 26.936m for BP-LANDMARC, and 20.243m for SVR-LANDMARC.

Most of the machine learning referred before only solve linear problem. There are deeper methods in machine learning to solve non-linear problem, called deep learning. Author in [14] defines deep learning as a technique that uses many non-linear information for execute either supervised or unsupervised of feature extraction, transformation, pattern analysis, and classification. There are two key aspects in deep learning: 1) The models consists of many non-linear information processing. 2) The methods for either supervised or unsupervised learning of feature extraction. Reasons that deep learning gains popularity in recent research are increased capability of GPU, lowered cost of computer hardware, and recent advances in signal processing.

Based on review of above papers, Bluetooth low energy has high chance become best candidate for indoor positioning, because low energy consumption that made devices usable longer, low cost in either installation or maintenance, has high update rates in receiving signal, and supported by modern smartphones. For estimation algorithm, regression using Deep Learning Convolutional Neural Networks (CNN) method is proposed to research if deep learning model is capable to increase positioning accuracy.

The proposed method is using BLE as signal transmitter and receiver. Current bluetooth technology (bluetooth 4.1) provides BLE with small, cost effectiveness, and lower energy consumption device that allows BLE to run for several years and designed for machine-to-machine communication. The optimized result of RSS signal is around radius 2-3 meters [15]. Fingerprinting is the state-of-the-art method. This method removes multipath effect and human movement problem from BLE radio signal weaknesses by collecting multiple samples and uses the average from samples [9]. A method of interval sampling also removed fast fading problem by increasing time interval to receive BLE RSS signal [3] [4].

The machine learning model used to predict estimated position is using CNN. The model estimates x position and y position separately which mean there are two models of machine learning with similar architecture. To train these models, RMSE is used as cost function. Results from these models are mean of distance and cumulative distribution function that represent the distances produced by these models.

III. Regression based Fingerprint Method

Design proposed in this paper is indoor positioning using BLE with deep learning CNN regression model for fingerprinting. Where The proposed method is based on fingerprinting that consist of two phases. Offline phase to create radio map database. A reference point will be assigned in the map. By standing at the point, the smartphone will receive RSS values from all of the BLE beacons. these RSS values will be stored together with the reference point as a single data in database. This step will be repeated until each reference point in the map has RSS values stored within the database.

Online phase estimate position using the CNN model. Position of a person will be estimated using localization algorithm. First, the smartphone receiver will receive RSS values within an interval of 5 seconds. The RSS values will be measured with Euclidean distance (2) with all reference points and then ranked using k-Nearest Neighbor to find five nearest reference points with estimated position.

There are two models of deep learning used to estimate x position and y position that trained separately. By using these distances, together with x position of the reference point related with the distance, feed them as inputs of the Deep learning machine. Resulting the estimated x position of the smartphone. These distances will be used again in estimating y position. The deep learning architecture in proposed method's illustration is represented in Fig. 1.

The model used to predict estimated position is using CNN. CNN used to process a grid-like data or matrix that can be applied with mathematical operation called convolution and modified the output using pooling function. Convolution operation is commutative way for giving weight to the measurement to provide a smoother measurement, resulting multidimensional array of data called feature map. Convolution involves three important ideas to improve machine learning. Sparse interactions using kernel smaller than the input, parameter sharing that uses same kernel parameter in more than one input functions. And equivariant representations that present the output changes in same way the input changes [16].

Pooling function replaces output at certain location with a summary statistic of nearby outputs. The purpose is to make representation of output approximately invariant to small translation of the input, so the feature placed exactly where it is unaffected by small transformation. Method of pooling used in this research is cross-channel pooling [17] to merges multiple feature maps into single feature map to reduce number of parameters needed.

The CNN architecture proposed is inspired by AlexNet [18] deep learning architecture as base principle as AlexNet is suited for small scale training datasets. Using 10x1 matrix contains RSS Euclidean distance and position (either x or y) from five best ranked reference points as input. The CNN starts with the input go through convolution by 3x1 kernel with one padding, so the output of convolution will not change either row or column of kernel matrix. Convolution is done using Rectified Linear Unit (ReLU) as activation function to produce five feature maps from a convolution. These results will go through cross-channels pooling layer to combine them into single 10x1 matrix. These processes are be called as convolution-pooling layer and modified to prove difference of accuracy of CNN model. This process will be repeated according to needs of research, resulting 10x1 matrix that will go through two fully connected layers. The fully connected layers are using ReLU method so the output can be either positive infinite or zero. Finally, the output layer is a neuron that produce 1 value of estimated position (either x or y). The CNN will be repeated for estimating another value of estimated position that has not through CNN process yet.

The fingerprinting method from the model will be trained using Root Mean Square Error (RMSE) [19], where real position will be subtracted by estimated position then find mean value from distance values, and square root the mean value. RMSE used for lost function, by comparing cost between two datasets (training and validation) from a model. The calculation will be done using (3) and Equation (4) Where (x, y) is real position and (xp, yp) is predicted position. Both trained models were compared with weighted sum or weighted centroid method to measure the performance using mean of Euclidean distance (5) and cumulative distributive function of estimated distances. The cumulative distribution is to find accuracy at certain percentile.

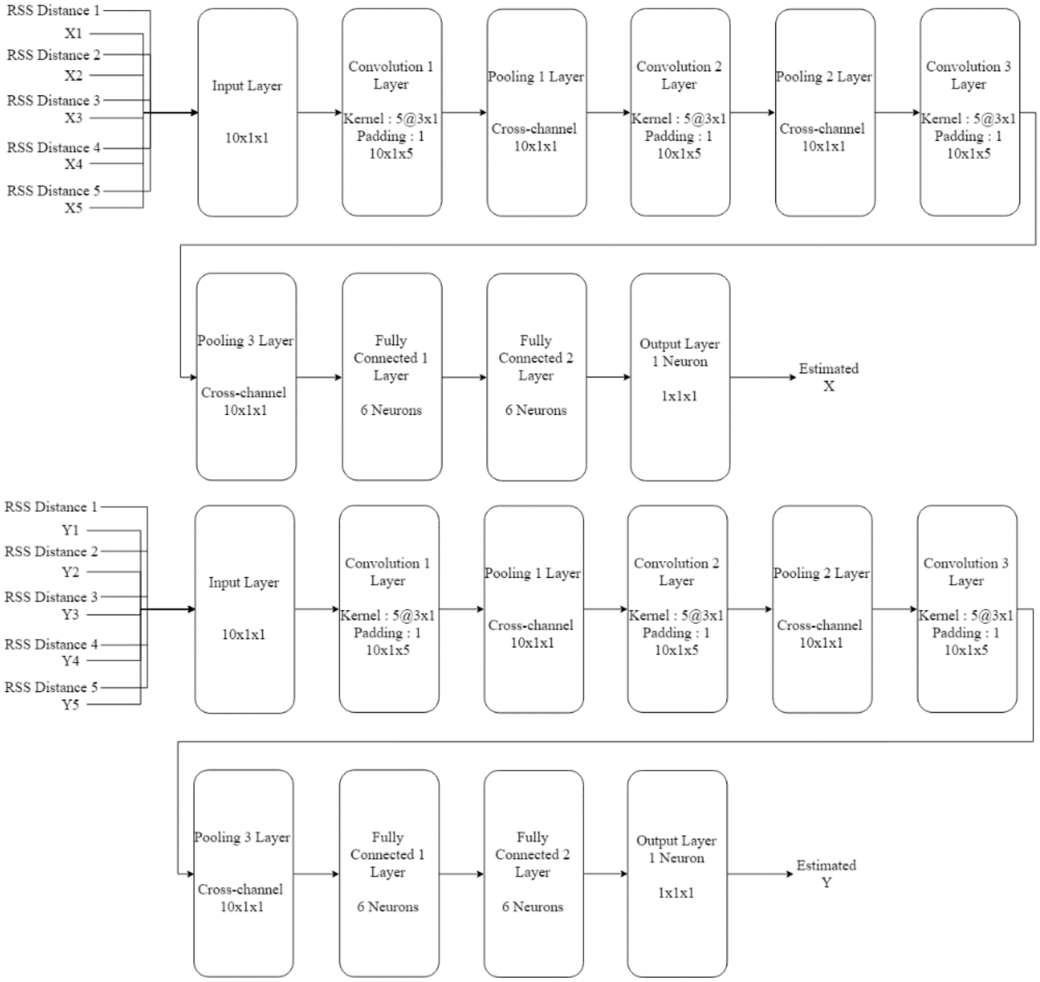


Fig. 1. CNN Model (with 3 Convolution-Pooling Layers).

IV. Experimental Design

BLE beacons used in this research are Nordic Semiconductor nRF51822 Bluetooth Smart Beacon that shown in Fig. 2. Each of them configured with 0 dBm transmit power and with 200 milliseconds of refresh rate. The beacons used highest possible settings to achieve maximum potential that can be predicted. 24 BLE beacons used in this research and each of them placed according to coordinate map based on their BLE ID. All BLE beacons attached at height around 1.2 meters.

Data used in the research are based on reference points and testing points in the coordinate map shown in Fig. 3. Sampling data are done in an office room with size of 12m x 19m with all office room properties. A total of 54 reference points used with 100 samples for each reference point and 156 testing points used with 10 samples for each of testing point. Reference points placed with gap of 2 meters and testing points placed with gap of 1 meter. Data sampling is done by standing on the point with a smartphone installed with RSS signal receiver application for few minutes. After all RSS values from each beacon received, the application starts to store the sample and repeat sampling with 1 second interval of each sampling to avoid zero value from a beacon. There are three sessions of data sampling separated in range of a week because of the limitation of building operational time, different amount of time to stabilize and receive all RSS signal from 24 beacons, and inferences by amount of people going around the moment of sampling.

Fig. 2. Nordic Semiconductor nRF51822 Bluetooth Smart Beacon.

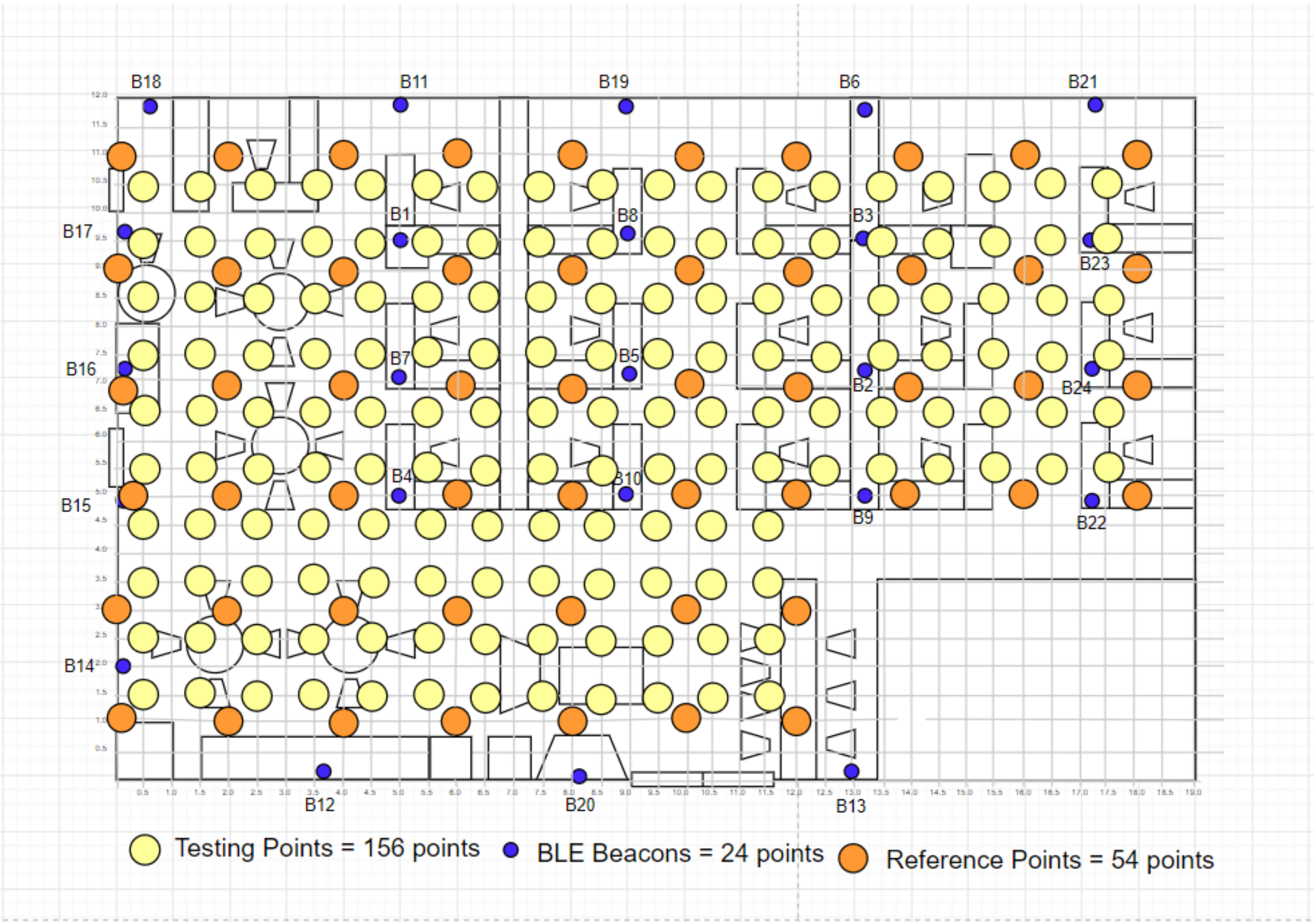


Fig. 3. Coordinate Map and BLE Beacons' Position.

Reference Points: RSS values from 100 samples will be averaged into single RSS for each beacon, producing 24 averaged RSS values for each reference point. Each RSS value from the reference point will be determined the distance with each testing points using Euclidean distance (2).

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(RSS_{ik} - RSS_{jk}\right)^2} \qquad (2)$$

This process results in 1560 data that ready to be used. The data separated into 1040 training data, 260 validation data, and 260 testing data by random sampling. Training data will be fed into the deep learning model (Convolutional Neural Network) for training and finally evaluated with validating data. Result from the training and validation will be measured using Root Mean Squared Error (RMSE) of predicted position with label position as the lost function. Both x and y RMSE value will be measured separately as (3) and (4).

$$RMSE_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[\left(x - x_p\right)\right]_i^2} \qquad (3)$$

$$RMSE_y = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[\left(y - y_p\right)\right]_i^2} \qquad (4)$$

Finally, by using testing data, the performance of the model is calculated using mean of distance (5) and cumulative distribution function (CDF). Using both combined result of distance from both x and y value predicted with their labels calculated using Pythagoras rule. The sum of combined result from each estimated position is divided by n-number of samples, producing the mean of distance.

$$Mean\ of\ Distance = \frac{1}{n}\sqrt{\sum_{i=1}^{n} d_i} \qquad (5)$$

CDF calculated using the sorted combined result of estimated position. Then, separate the value by ranking them into n-percentile. The percentiles are based on sample rank position divided by n-number of samples. Which mean lowest percentile is called minimum distance and highest percentile is called maximum distance. To clarify the results, a set of minimum distance, median distance, 90th percentile distance, and maximum distance are used. Results from research are shown in Fig. 4.

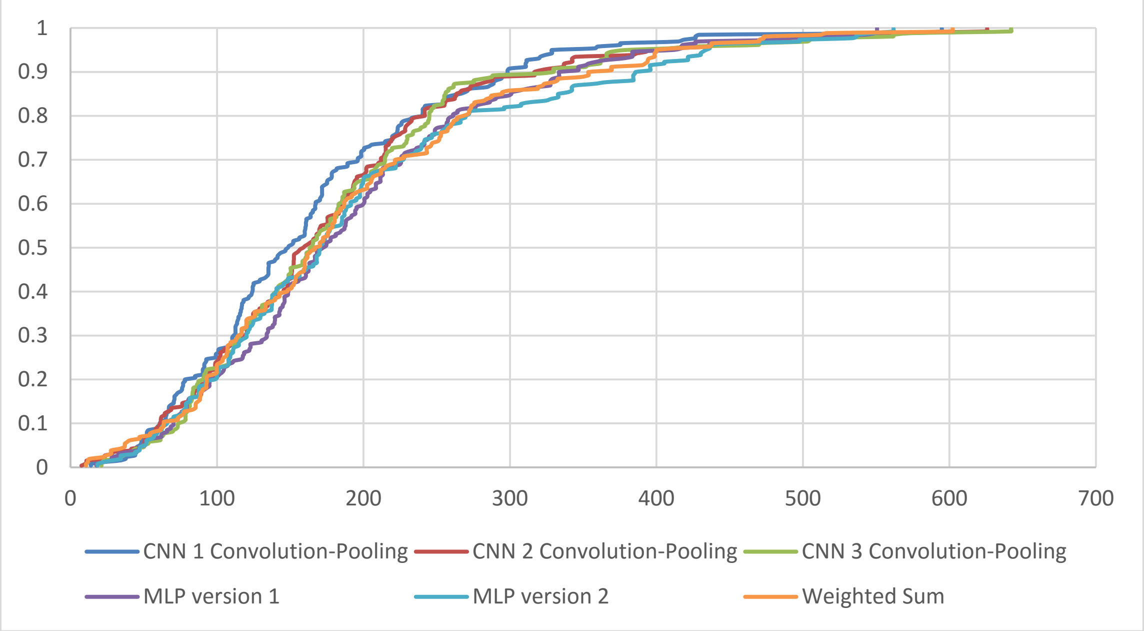


Fig. 4. Cumulative Distributive Function for Distance Accuracies.

V. Result

The experiment conducted using 1560 data collected. The data separated as described in previous section and randomized for each training epoch using random sampling, with exception for testing data. These data used to feed both CNN models, Multilayer Perceptron (MLP) models, and weighted sum method. Few variations of models are tried to prove accuracy of the method. The first CNN model used 1 convolution-pooling layer; second CNN model used 2 convolution-pooling layers; and the third one using 3 convolution-pooling layers. Then after the convolution-pooling process, the result went through 2 fully-connected layers with 6 neurons each fully-connected layer. For the MLP models, 2 variations are used. The first one is using 2 layers with 6 neurons each. This one is similar with CNN models but without convolution-pooling layers. The second one is using 6 neurons on first layer, then 3 neurons on second layer.

The training and validation result for CNN models shown on Fig. 5 and Fig. 6, while MLP models shown on Fig. 7 and Fig. 8. The RMSE cost results from last epoch of model x are 137.13cm for training and 127.92cm for validate. For model y, it was 163.39cm for training and 156.67cm for validate. For MLP Models, the RMSE cost results for model x are 150.59cm for training and 161.42cm for validate. The model y gave 181.61cm for training and 171.67cm for validate. Based from RMSE results, the CNN models give better cost reduction compared with MLP. Then, the models were measured for performance using 260 random testing data. Both value x and value y of position combined to calculate the distance. Results from comparing mean of distance are shown in Table II, which consisting of 167.49cm from first CNN model, 179.96cm for second CNN model, and 183.40cm for third CNN model. The MLP first model gave 192.22cm and second model gave 195.95cm. The weighted sum gave 189.50cm. From mean of distances, the CNN models gave better results compared with MLP models and weighted sum.

From the training and validate results of CNN and MLP, CNN model shown faster learning rate on both model x and model y compared with MLP model.

Cumulative distributive function from both model’s distances are shown in Fig. 4. First model of CNN gave 298.36cm at 90th percentile. This model gave the highest performance compared with other models. Second model of CNN gave 316.69cm at 90th percentile, while third model gave 329.81cm at 90th percentile. The MLP models did not outperform the CNN models. MLP first model gave 333.50cm at 90th percentile and second model gave 385.79cm at 90th percentile. The benchmark weighted sum model gave 353.57cm at 90th percentile, which in between the MLP models.

The experiment shows that CNN model achieved accuracies of < 330cm at 90th percentile. Weighted sum model achieved accuracies of < 360cm at 90th percentile. The MLP models are in between the benchmark method but could not outperform CNN models. In Fig. 4, it was shown that CNN models performed slightly better than MLP models and weighted sum. Table III shown cumulative distribution for certain percentile from all models used.

TABLE II. Mean of Distance in CM

Model	**Mean of Distance**
CNN model 1 Convolution-Pooling	167.49
CNN model 2 Convolution-Pooling	179.96
CNN model 3 Convolution-Pooling	183.40
MLP Version 1	192.22
MLP Version 2	195.95
Weighted Sum	189.50

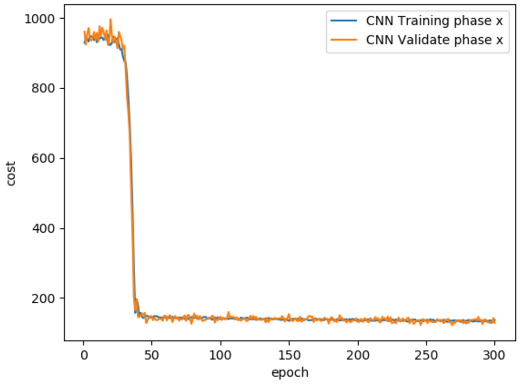


Fig. 5. Training and Validate for CNN Model x.

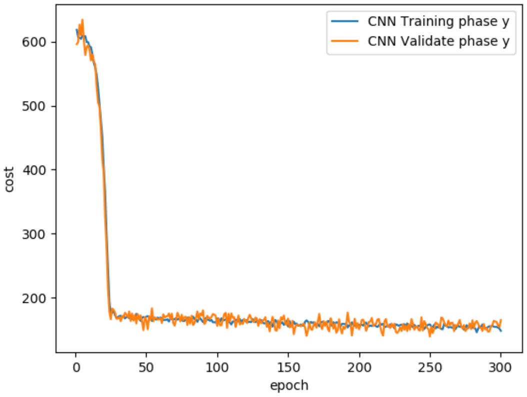


Fig. 6. Training and Validate for CNN Model y.

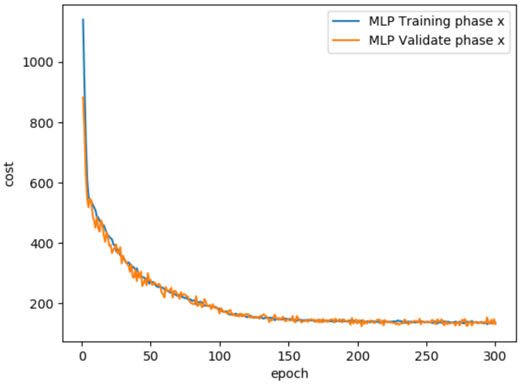


Fig. 7. Training and Validate for MLP Model x.

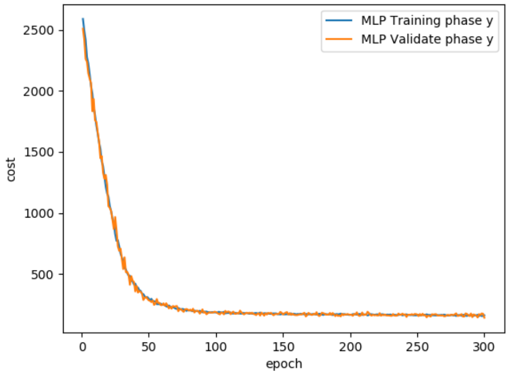


Fig. 8. Training and Validate for MLP Model y.

TABLE III. CUMULATIVE DISTRIBUTION OF DISTANCE IN CM

Percentile	**Min distance**	**Median distance**	**90th percentile distance**	**Max distance**
CNN model 1 Convolution-Pooling	13.93	148.41	298.36	594.79
CNN model 2 Convolution-Pooling	7.414	158.85	316.69	625.73
CNN model 3 Convolution-Pooling	21.11	163.38	329.81	642.11
MLP Version 1	17.69	172.40	333.50	550.57
MLP Version 2	18.44	170.21	385.79	561.83
Weighted Sum	10.60	166.77	353.57	602.20

VI. DISCUSSION

With the limitation of data collection time, the amount of data used in this research is pretty low for a CNN model. The proposed method might be not the most optimal model designed. The model needs to be analyzed and optimized with the principle of Deep Neural Network (DNN) [20] to further increase the accuracy.

Another problem in data collection is the battery capacity. As it is using button battery, the BLE could stay active at least six months to two years with relatively stable signal power [21]. However, the BLEs used in this research can only stay active for not more than one month. This means that current BLE beacons' setting is using too much battery power.

This research used averaging method to solve unstable RSS values received from each BLE. There are large fluctuations of data received caused by large amounts of people and obstacles in the office room. The placement of BLE beacons could affect the positioning error resulted from prediction [22]. More localization model could also be tried to improve the quality of RSS values. As this model affect the coverage area of the BLE beacons placed in the office room [23].

VII. CONCLUSION

This paper proposed CNN architecture to estimate position. The experiment showed that proposed CNN model surpassed MLP models and weighted sum method. The CNN models gave accuracies of < 330cm at 90th percentile, while the weighted sum gave accuracies of < 360cm at 90th percentile while MLP models in between CNN models and weighted sum. However, the CNN model has not been modified with optimum configuration and not yet implemented with different environments. This research is executed with maximum capability from BLE beacons used, which is significantly reducing BLEs' battery lifetime. The extensive time of sampling data should be reduced to improve accuracies.

VIII. FUTURE RESEARCH

For future researches, different CNN architectures should be tested and compared. These models should be tested on different indoor environments and room shapes. Another point to be researched is the validity of CNN model for predicting position. More researches should be done to prove it.

Different BLE placements could also be interesting future topic to analyze changes. The optimum setting for BLE beacons used in this research is not yet to be defined as the current setting gave high battery power usage but provide maximum capabilities from the BLE beacon.

REFERENCES

[1] R. F. Brena, J. P. García-Vázquez, C. E. Galván-Tejada, D. Muñoz-Rodriguez, C. Vargas-Rosales and J. Fangmeyer, "Evolution of indoor positioning technologies: A survey.," Journal of Sensors,2017, 2017.

[2] D. Čabarkapa, I. Grujić and P. Pavlović, "Comparative analysis of the bluetooth low-energy indoor positioning systems," in 2015 12th International Conference on Telecommunication in Modern Satellite, Cable and Broadcasting Services (TELSIKS), 2015.

[3] R. Faragher and R. Harle, "An analysis of the accuracy of bluetooth low energy for indoor positioning applications," in Proceedings of the 27th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GNSS+'14), 2014.

[4] R. Faragher and R. Harle, "Location fingerprinting with bluetooth low energy beacons," IEEE journal on Selected Areas in Communications, vol. 33, no. 11, pp. 2418-2428, 2015.

[5] M. Terán, J. Aranda, H. Carrillo, D. Mendez and C. Parra, "IoT-based system for indoor location using bluetooth low energy," in Communications and Computing (COLCOM), 2017 IEEE Colombian Conference, 2017.

[6] F. Topak, M. K. Pekeriçli and A. M. Tanyer, "Technological Viability Assessment of Bluetooth Low Energy Technology for Indoor Localization," Journal of Computing in Civil Engineering, vol. 32, no. 5, p. 04018034, 2018.

[7] D. Contreras, M. Castro and D. S. de la Torre, "Performance evaluation of bluetooth low energy in indoor positioning systems," Transactions on Emerging Telecommunications Technologies, vol. 28, no. 1, p. e2864, 2017.

[8] Y. Zhuang, J. Yang, Y. Li, L. Qi and N. El-Sheimy, "Smartphone-based indoor localization with bluetooth low energy beacons," Sensors, vol. 16, no. 5, p. 596, 2016.

[9] S. He and S. H. G. Chan, "Wi-Fi fingerprint-based indoor positioning: Recent advances and comparisons," IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 466-490, 2016.

[10] F. Yu, M. Jiang, J. Liang, X. Qin, M. Hu, T. Peng and X. Hu, "5 G WiFi Signal-Based Indoor Localization System Using Cluster k-Nearest Neighbor Algorithm," International journal of distributed sensor networks, vol. 10, no. 12, p. 247525, 2014.

[11] C. Li, Z. Qiu and C. Liu, "An improved weighted k-nearest neighbor algorithm for indoor positioning," Wireless Personal Communications, vol. 96, no. 2, pp. 2239-2251, 2017.

[12] S. Tuncer and T. Tuncer, "Indoor localization with bluetooth technology using artificial neural networks," in Intelligent Engineering Systems (INES), 2015 IEEE 19th International Conference, 2015.

[13] H. Xu, M. Wu, P. Li, F. Zhu and R. Wang, "An RFID Indoor Positioning Algorithm Based on Support Vector Regression," Sensors, vol. 18, no. 5, p. 1504, 2018.

[14] L. Deng and D. Yu, "Deep learning: methods and applications," Foundations and Trends® in Signal Processing, vol. 7, no. 3-4, pp. 197-387, 2014.

[15] K. Townsend, C. Cufí and R. Davidson, Getting started with Bluetooth low energy: tools and techniques for low-power networking, O'Reilly Media, Inc., 2014.

[16] I. Goodfellow, Y. Bengio, A. Courville and Y. Bengio, Deep learning, vol. 1, Cambridge: MIT press., 2016.

[17] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville and Y. Bengio, "Maxout networks," in arXiv preprint arXiv:1302.4389, 2013.

[18] S. Hoo-Chang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE transactions on medical imaging, vol. 35, no. 5, p. 1285, 2016.

[19] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," Geoscientific model development, vol. 7, no. 3, pp. 1247-1250, 2014.

[20] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*, Jerusalem, 2015.

[21] Y. Wang, Q. Yang, G. Zhang and P. Zhang, "Indoor positioning system using Euclidean distance correction algorithm with bluetooth low energy beacon," in *2016 International Conference on Internet of Things and Applications (IOTA)*, Pune, 2016.

[22] W. Nakai, Y. Kawahama and R. Katsuma, "Reducing error of positioning based on unstable rssi of short range communication," in *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, Krakow, 2018.

[23] C. Liu, D. Fang, Z. Yang, H. Jiang, X. Chen, W. Wang, T. Xing and L. Cai, "RSS distribution-based passive localization and its application in sensor networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2883-2895, 2015.

A New Security Model for Web Browser Local Storage

Thamer Al-Rousan[1]
Faculty of Information Technology
Isra University
Amman-Jordan

Bassam Al-Shargabi[2]
Faculty of Information Technology
Middle East University
Amman-Jordan

Hasan Abualese[3]
Faculty of Information Technology
Ajloun National University
Ajloun-Jordan

***Abstract*—In recent years, the web browser has taken over many roles of the traditional operating system, such as acting as a host platform for web applications. Web browser storage, where the web applications can save data locally was one of the new functionalities added in HTML5. However, web functionality has increased significantly since HTML5 was introduced. As web functionality increased, so did the threats facing web users. One of the most prevalent threats was the user's privacy violations. This study examines the existing security issues related to the usage of web browser storage and proposes a new model to secure the data saved in the browser's storage. The model was designed and implemented as a web browser extension to secure the saved data. The model was experimentally demonstrated and the result was evaluated.**

***Keywords*—HTML5; security; local storage**

I. Introduction

The internet and its applications have critically influenced us and becoming the main part of daily life. The internet provides information and opportunities, which were not accessible previously. The digital population has risen in the last few years; it reached 4.087 billion users in April 2018 [1]. With an increase in dependency on the internet, the importance of privacy protection has become significant.

When browsing, users expose themselves and their personal information to several threats, such as malware or phishing, which directly disrupt privacy [2]. A user's profile has been the target of many attacks. There are different reasons for an attack on a user's profile, including identifying the characteristics of users to meet their requirements or selling this information for surveillance or advertising [3]. Over the years, many technologies have been utilized to track a user's profile; the most well-known is cookies.

HTML is a standard language used to develop web applications. HTML5 is the latest version of HTML that the W3C standard introduced as a written language for web pages and associated applications (APIs). HTML5 brings a set of new advancements to the browser, which didn't exist previously, including enhanced XMLHttpRequest (XHR) and webSocket [4]. Web-based applications can be run offline via local storage, meaning the data will be saved in the user-side storage and it can be accessed without having to connect to the network [5]. User -side storage as indexed database APIs, web SQL databases, and web storage revolutionized the web and minimized the differences between native applications and web-based applications. Consequently, network connections' lack of effectivity on the server-side and the visual uncertainty affected by refreshes could be avoided. Besides, HTML5 offers functionality that can be utilized by web-based applications on any platform and on different browsers. Thus, modern web browsers support the plurality of new HTML5 features [6].

In spite of the advantages, user-side storage did not come without a price. The data saved by user-side storage is unencrypted, so users may experience privacy violations, such as a tracking vector. The literature in [7], [8], [9] revealed that the data created by web applications and saved in user-side storage cannot be completely deleted, even when the users tried. Thus, there are dangerous privacy implications, such as the details of a user's bank account or credit cards.

Previous studies to protect user-side storage were rather limited. This study proposes a new model that aims to protect the user's privacy. The proposed security model, which was based on the JavaScript encryption library (JSEL), was implemented as a web browser extension. The browser extension offered complete security protection, as the data were saved in encrypted form.

The remainder of the paper is arranged as follows. The theoretical background is presented in Section 2. Section 3 presents an overview of possible technical attacks associated with local browser storage and the current security solutions. Section 4 presents a new security model for browser-based storage. Related works are discussed in Section 5. The experimental study and the evaluation are presented in Sections 6 and 7, respectively. Finally, the conclusions and future work are presented in Section 8.

II. Background

The most famous user-side database is HTTP cookies. Cookies are small packets of data (normally 4KB) sent from websites servers to users through HTTP headers or by utilizing user-side scripting. Cookies saved in a user's browser are on the user's machine. Every cookie is related to a source, i.e., a port number, HTTP protocols, and the hostname. It based on a security mechanism termed same-origin policy (SOP). Website servers use cookies to remember a user's stateful information or to trace the browsing activity of the user. Security and privacy are the major challenges, as an attacker can steal data from cookies [10].

Over the years, many technologies appeared to save structured data in user-side storage. The majority appeared through third-party plugins, such as Oracle Java, Object of Adobe Flash, Google Gears, and Microsoft Silverlight [7]. As HTML5 appeared, browsers started to replace third-party plugins with built-in functionalities and new user-side storage technologies were established, such as "indexed database API, web SQL database, and web storage" [11].

Web storage provides a way for web applications to save data locally in the user's browser without affecting website performance and provides more storage capacity than cookies. Depending on the user's browser, web storage capacity can be anywhere from 5MB to 25MB [11]. Web storage saves data in key/value pairs in the browser, and there two types. First is session storage data, which will be lost when the user browser is closed. Second is local storage data that are maintained even when the browser is closed [12]. Web storage is dependent on user-side scripting, such that the web application can retrieve locally stored data by utilizing a user-side JavaScript API even when the user's browser is disconnected [11]. Besides, web storage differs from cookies, since the saved data cannot be transferred through HTTP headers. The security policy of web storage is the same as for cookies since each source is assigned to a unique web storage object. Therefore, using web storage on websites that do not support HTTPS or that use hostname sharing is not recommended [6].

A WebSQL database is another way for web applications to save huge amounts of information in the user's browser, which can be queried by using standard SQL syntax. A WebSQL database is similar to Google Gears, and both depend on SQLite [13]. The main weakness of a WebSQL database is that it is not supported by W3C, and many browsers, for example, Mozilla, have decided to stop supporting WebSQL databases. Despite the W3C decision, three main browser vendors (Opera, Safari, and Chrome) still support WebSQL databases [11].

Indexed database API (Indexed DB) is a transactional database system that came from the W3C specification in 2009. Indexed DB is a substitute for the deprecated WebSQL database [14]. Indexed DB has the same accessing mechanism as web storage, but the scale and structure of the saved data are different. Web storage saves data in key/value pairs and is suitable when the dataset is simple, whereas Indexed DB allows for a large amount of structured data to be saved in storage [15].

Compared to SQL-based relational database management system (RDBMS), which uses tables to save data, Indexed DB is an object-oriented database. Indexed DB allows for objects to be saved/retrieved with keys [16]. Web applications can only access and operate saved structured data through the API provided by Indexed DB. Indexed DB is built on a transactional database mode and is typically key-value asynchronous storage. It provides rapid access to a lot of organized information [14]. In contrast, the security mechanism for Indexed DB is no different than web storage. The Indexed DB security mechanism is based on the principles of the SOP [13].

III. Security Issues

With the new functionalities of HTML5, which increased the accessibility to a user's computer resources, including offline caching and local storage, new security issues arose. Most web browser security methods were not upgraded to the new technologies of HTML5. The only browser security method in place against possible risks was SOP [9]. SOP links the saved information to a specific domain, so the information can only be retrieved from the original domain. When SOP is applied, the browser tests the port number, name of host, and protocol against the source record of the saved data [2].

HTML5 has numerous new modules, such as "XMLHttpRequest (XHR), the document object model (DOM), and cross-origin resource sharing (CORS)." HTML5 brought new technologies, such as local storage, web Socket, and enhanced XHR. With these technologies and modules, it has improved the surface of the attack and added new risks to the end-user, which means that the SOP will not help [2]. The weaknesses of SOP led to many aggressions, such as "cross-site scripting (XSS), cross-site request forgeries (CSRF), cross-origin resource sharing (CORS) attacks, social engineering, and physical access" [17].

CORS is a technique that lets JavaScript make an XHR request from other domains outside the original domain. XHR is used by JavaScript to perform transfers between the server and the user. Cross-domain requests are normally prevented by web browsers, so CORS adds additional information to HTTP headers to permit the request [17]. A CORS technique allows for several domains (cross-domain requests) between the user and the server. However, hackers can make cross-site request forgeries by bypassing SOP and creating "cross-domain connections" to permit the deployment of a CORS attack vector [3].

XSS is one the most common attacks in a web-based application. Based on the open software security community (OWASP) list, XSS ranks third place in OWASP list [18]. XSS is a security weakness, where malicious code (generally JavaScript) is injected into a web-based application and is comprised of dynamic content sent to a victim's browser. The victim's browser cannot identify that the script is infected, and will execute the malicious script without being validated because it appears to have arrived from a trusted source. The malicious script can change or transmit session tokens, cookies, or any information via the victim's browser [19]. There are many techniques to prevent XSS attacks, such as encoding output (stop executing URL links that include binary encoded characters), filtering (the input data is filtered before it saved in the database), and authenticating user input (checking the format of the user) [6].

Social engineering is a way of fooling and cheating users so that personal data is shared, such as bank accounts or passwords. The hackers can install malicious code to access data by controlling the user's computer. Social engineering techniques are easier than other hacking techniques and can include calling by phone, sending an email, and real-life chats. An email from a known sender is a popular example of a social engineering technique. Hackers send an email message with a link to the victim contact list members. The link may contain a picture, audio, movie, or file that the malicious software is inserted in. If the member clicks on this link, that member could then be the next victim [20].

CSRF is also known as session riding or a one-click attack. An attacker could trick the victim into unintentionally performing an undesirable action in the web application [21]. CSRF is normally performed with the assistance of social engineering techniques. As the victim is currently authenticated by the web application, it is difficult for the server to differentiate between legal requests and fake requests [15]. CSRF attacks can harm both users and businesses. It can result in damaged user relationships, illegal money transfers, changed passwords, and stealing of data, including the session cookies [17].

Physical access is another security problem that occurs when the attacker has access to the user's computer. When this happens, the saved data and even the deleted data can be stolen. Recovering deleted data is primarily concerned for developers. The experiment conducted by Shahryar [22] showed that the location of the deleted data within a database was still physically reserved. The deleted data was only marked as deleted ,but the data existed in storage and could be retrieved with forensic tools. In most web browsers, the non-deleted data will not be overwritten by newly saved data [9]. The data is saved in storage as non-encrypted, so it is not protected and presents a possible security issue. Encryption will prevent data from being attacked. User-side encryption is important since it offers a security mechanism for saved data and blocks unauthorized persons from stealing the user's data. Unfortunately, user-side encryption is not yet mature. In the next suction, this study proposes a new encryption model to protect a user's data.

IV. Security Model

This study proposes a new security model to address issues caused by saving data in an unsafe manner. In the proposed model, an additional layer was added among the web browser storage and the web browser itself. The model contains a new algorithmic framework for improving security against threats. The model uses JavaScript encryption library, which was applied to the web browser as “an extension.” The extension was located at the top of the web browser storage, such that whenever writing or reading information, it is encrypted.

Encryption/decryption can be implemented inside browsers via the” JavaScript crypto library.” There are a small number of JavaScript encryption libraries that may be applied to encrypt data in the client’s browser, such as WebCryptoAPI, PolyCrypt, Cryptos, Jscrypto, and SJCL [21]. Many factors affect the selected encryption library, including library size, hashing functions, key lengths, keyed-hash message authentication codes (HMAC), salt, and availability of multi-platform and multi-browser options [11].

This study selected the Stanford JavaScript crypto library (JSCL) to implement the encryption/decryption on the user’s browser. The SJCL library was chosen because it is a secure, small, multi-platform, and powerful library [23]. SJCL uses the advanced encryption standard (AES) to encrypt data at key sizes of 128/192/256 bits. It also uses an HMAC validation code, the SHA256 hash function, OCB and CCM authenticated modes, and the PBKDF2 to strengthen the password. BKDF2 applies a pseudorandom function to the entered password together with a salt value and reiterates the process several times to generate a derived key, which then can be utilized as a crypto key. PBKDF2 verifies each message it sends to avoid any changes. Furthermore, the tests conducted in several browsers on Windows, Linux, and Mac have shown that the SJCL was faster than any existing encryption library [23].

As the browser extension used the SJCL library, the same shared key was applied for both encryption and decryption, which made the data prone to malicious attacks. To overcome the limitations of the SJCL library, the study combined the RSA standard [9] with SJCL library. The combination resulted in a hybrid encryption algorithm comprised of both RSA and AES to guarantee data integrity.

When the browser storage received a request from a web application to save data, the data was encrypted by the proposed browser extension. This ensured that the data was safe and secured against illegal access even if the attacker gained physical access to the machine. The data encryption/decryption steps are as follows:

- Get Login: The initial step is to afford a secure log into the system. Web applications can handle this step for the users by using the login process.
- Data Encryption: When a new request from a web application arises to save data in the database, the plain text data (PTORG) is encrypted by the designed extension and a public key (KSJCL) is generated. The KSJCL will be used when the SJCL library encrypting the data. The encrypted data is named ENORG. Simultaneously, when the encryption process is enabled in user-side, the RSA standard is generated. RSA standard is used to encrypt the ENORG by using the private key (PSJCL) to produce a “digital signature.”
- Data Decryption: When a request from an authorized user to read data from the browser’s storage arises, the key (K SJCL) is encrypted using an RSA standard with the requester public key (RKPB). The encrypted key is named (RKEN SJCL). An XML file is then created, including ENORG and RKEN SJCL. Finally, the ENORG is decrypted using KSJCL.
- Deletion of Data: The data in the session storage is safely deleted - that is, it was replaced with zeros. Thus, the deleted data cannot be read again, since the original data was set to zero.

The goal of this study is to protect saved data on client-side databases against illegitimate access. The study had many choices for encryption, but most tended to save encryption keys on the server-side and not on the user-side. One limitation of utilizing the server-side was associated with offline storage, the client won't almost certainly get to the information. Besides, if the webserver is vulnerable to any attack, then the encryption key will be perilous and thus, the encrypted information will be defenseless against unapproved access from attackers. Also, the W3C recommendation is to save the encryption keys on the client-side [24].

Browser security models depend on an SOP mechanism. However, an SOP mechanism alone is not enough for sophisticated web-based applications' local storage security. The proposed model differs in the security mechanism. Where the browser security models attempt to secure data amongst the user's browser and web-based applications, the proposed model secures data saved in the user's database. Thus, users can visit other websites without worrying about the databases.

V. Related Work

Previous studies regarding securing web browser storage and its effectiveness are still in early stages and are limited. Aggarwal et al. [25] conducted one of the first studies to analyze browsing security vulnerabilities. The results revealed that there was an insufficient implementation of the security mechanism in different web browsers, which pointed to user activities. Additionally, security control for Firefox was proposed, which protected users after private mode was enabled. In 2011, Oh et al. [26] analyzed the log files generated by a web browser, concentrating on search history, timeline analysis, and URL encoding. A Classification of Web Browser Log (CWB) tool was proposed to prove the analysis. Unfortunately, in the experiments, old versions of browsers were used that are currently outdated.

Ohana and Shashidhar [27]studied portable browsers to determine if data still existed after the browser session stopped. Similarly, Said et al. [28] analyzed the RAM in the browser sessions, including authorizations, history, images, and videos. Satvat et al. [29] extended the work by analyzing the file system and the network, which exposed significant contradictions in the browsing implementation that violated a client's privacy. Heule et al. [30]developed a security access control to protect confidential data that could be accessed and used by attackers, focusing on JavaScript extensions in Chrome. In the same manner, Lerner et al. [31] studied JavaScript extensions in Firefox from different perspectives, such as social, safety, and debugging, to determine which may be malicious.

Ruiz et al. [32] concentrated on recovery techniques created during browsing. Experiments within four personal levels showed how the browser could be stopped, namely shutdown, power down, freeze, and kill processes. The results revealed that all levels included user privacy violations in terms of obtaining browsing data. In the same manner, Montasari and Peltola [33] studied both file systems and RAM in different browsers. The results revealed that the most secure browser was Chrome and in second place was Firefox.

A survey conducted by Gao et al. [34] concentrated on awareness from a user's viewpoint. Authors surveyed more than 200 users regarding the security and privacy mechanisms provided by the most common web browsers on smartphone and desktop platforms. The results revealed that better security guarantees were required concerning a user's privacy. Tsalis et al. [35] studied the protection provided by different popular web browsers. A set of web data created in a usual browsing session was used to determine where these data were saved after the session was stopped. The results revealed that the deleted data after the browsing session were discovered straightforward or in a roundabout way in the database. Subsequently, any person who had "physical access" to the user's machine with adequate IT abilities could access these browsing data. In addition, nearly all browsers provided a similar protection level, and only Chrome in the guest mode provided better protection. Belloro and Mylonas [30] investigated the most common user storage methods, namely "Web SQL database, web storage, and indexed database." The outcomes revealed that web storage was the most utilized. Belloro and Mylonas also surveyed whether well-known mobile and desktop browsers that used "indexed database API, web SQL database and web storage" could defend clients from privacy violations. The results showed that the maturity of the security controls was inadequate to avoid privacy violations.

VI. Experiment

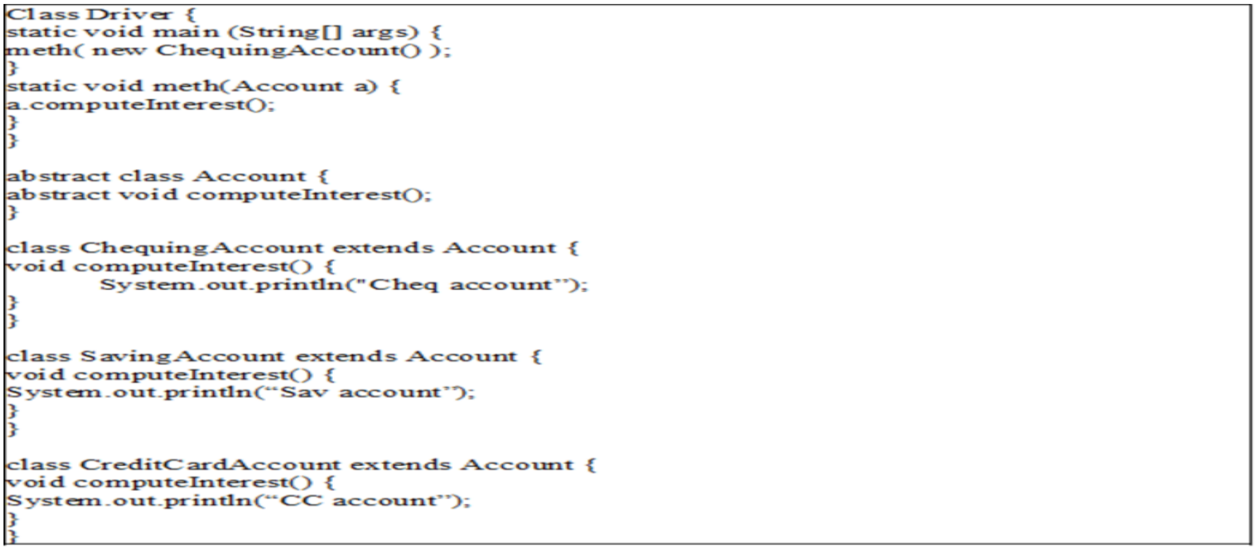
The experiment performed in this study was based on performance. The study tested the speed of encryption/decryption, which affected the performance of the model. Thus, the study tested the performance on the user-side to prevent the network and server latency impacting the results. The study selected indexed database as a representative for web browser local storage.

Before running the test, the code was executed for a compatibility test in diverse browsers, including Firefox, Google Chrome, Safari, Mozilla, and Opera. The latest Google Chrome browser hosted by a Windows10 server was chosen as the test environment. The server was an HP with Intel I7 8550U, and 8 GB DDR4 RAM. The test executed a small JavaScript by dexie.js, which is a powerful library for an indexed database and it can accelerate performance. Google private mode was used to the grantee that the test was executed without any additional load of scripts or any extension. Fig. 1 show the code used.

The test was performed in steps. The first step was to save the data without encryption (in a standard manner). The second step was to save data with encryption, as seen in Fig. 2. The same password and username were saved with encryption.

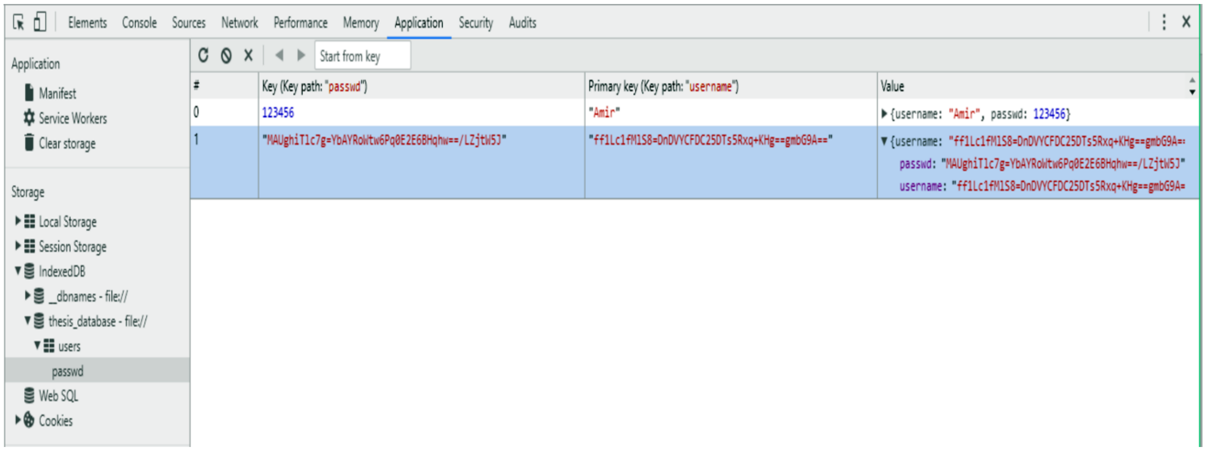
The test was executed without any user intervention. JavaScript read the entries of the database, placed the data into the table, and read it as a console output (console.log). All references employed in this study were based on Google performance analysis [17]. Fig. 3 shows the results of the performance test, it shows the top 10 bottom-up of processes that most consuming time.

Because the indexed database is a NoSQL database, the script run time had to be tested independently of the transaction speed of the database. Consequently, the performance test depended on the reading and writing of database entries in a loop. The time was measured before the loop started and after the loop was finished. Finally, the data entries contained "Password + I and User + i" where "I" was the loop counter. The time in milliseconds to insert database entries with and without encryption in Google Chrome is shown in Table I.



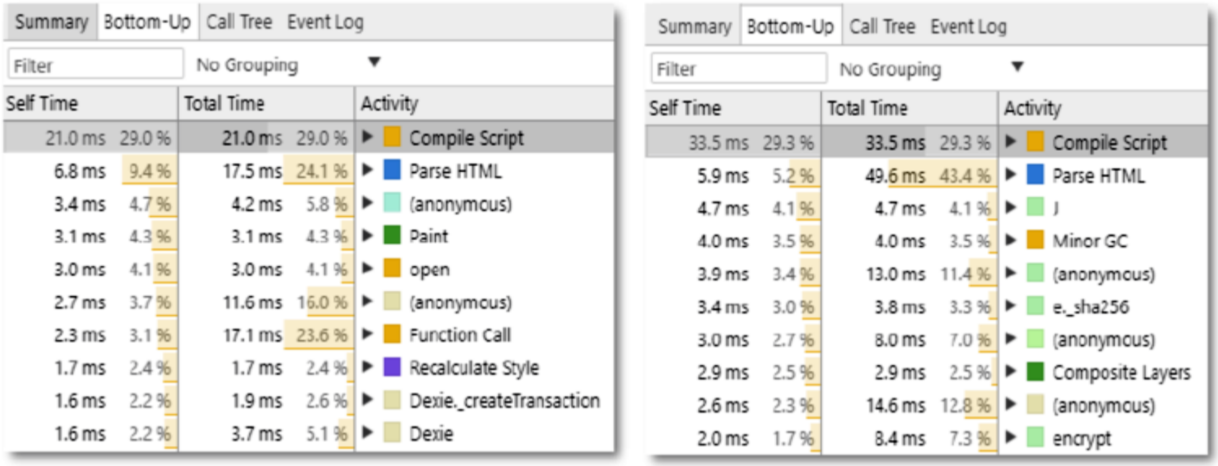
```
Class Driver {
static void main (String[] args) {
meth( new ChequingAccount() );
}
static void meth(Account a) {
a.computeInterest();
}
}

abstract class Account {
abstract void computeInterest();
}

class ChequingAccount extends Account {
void computeInterest() {
        System.out.println("Cheq account");
}
}

class SavingAccount extends Account {
void computeInterest() {
System.out.println("Sav account");
}
}

class CreditCardAccount extends Account {
void computeInterest() {
System.out.println("CC account");
}
}
```

Fig. 1. AES Sample.



#	Key (Key path: "passwd")	Primary key (Key path: "username")	Value
0	123456	"Amir"	▶{username: "Amir", passwd: 123456}
1	"MAUghiTlc7g=YbAYRoWtw6Pq0E2E6BHqhw==/LZjtW5J"	"ff1Lc1fM1S8=DnDVYCFDC25DTs5Rxq+KHg==gmbG9A=="	▼{username: "ff1Lc1fM1S8=DnDVYCFDC25DTs5Rxq+KHg==gmbG9A== passwd: "MAUghiTlc7g=YbAYRoWtw6Pq0E2E6BHqhw==/LZjtW5J" username: "ff1Lc1fM1S8=DnDVYCFDC25DTs5Rxq+KHg==gmbG9A=

Fig. 2. Key and Values after Plain Text Encryption.



Self Time		Total Time		Activity
21.0 ms	29.0 %	21.0 ms	29.0 %	▶ Compile Script
6.8 ms	9.4 %	17.5 ms	24.1 %	▶ Parse HTML
3.4 ms	4.7 %	4.2 ms	5.8 %	▶ (anonymous)
3.1 ms	4.3 %	3.1 ms	4.3 %	▶ Paint
3.0 ms	4.1 %	3.0 ms	4.1 %	▶ open
2.7 ms	3.7 %	11.6 ms	16.0 %	▶ (anonymous)
2.3 ms	3.1 %	17.1 ms	23.6 %	▶ Function Call
1.7 ms	2.4 %	1.7 ms	2.4 %	▶ Recalculate Style
1.6 ms	2.2 %	1.9 ms	2.6 %	▶ Dexie._createTransaction
1.6 ms	2.2 %	3.7 ms	5.1 %	▶ Dexie

Self Time		Total Time		Activity
33.5 ms	29.3 %	33.5 ms	29.3 %	▶ Compile Script
5.9 ms	5.2 %	49.6 ms	43.4 %	▶ Parse HTML
4.7 ms	4.1 %	4.7 ms	4.1 %	▶ J
4.0 ms	3.5 %	4.0 ms	3.5 %	▶ Minor GC
3.9 ms	3.4 %	13.0 ms	11.4 %	▶ (anonymous)
3.4 ms	3.0 %	3.8 ms	3.3 %	▶ e._sha256
3.0 ms	2.7 %	8.0 ms	7.0 %	▶ (anonymous)
2.9 ms	2.5 %	2.9 ms	2.5 %	▶ Composite Layers
2.6 ms	2.3 %	14.6 ms	12.8 %	▶ (anonymous)
2.0 ms	1.7 %	8.4 ms	7.3 %	▶ encrypt

Fig. 3. The Results in the Performance Test; Left the Plain Text Entries, Right Encrypted Entries.

TABLE. I. THE TIME TO INSERT DATA INTO INDEXEDDB

Test Number	Record Number	Without Encryption (ms)	With Encryption (ms)
1	1	2	5
2	100	5	53
3	500	14	109
4	1000	21	202
5	5000	87	515
6	10000	126	1185
7	50000	554	4760
8	100000	956	9008
9	500000	6350	66187

VII. EVALUATION

Big O notation was used to evaluate the proposed model. It compares the effectiveness of different algorithms by revealing the time that the algorithm took to run. The runtime can be expressed with Big O Notation as how fast the runtime grew relative to the size of the input "n" (denoted O (n)).

AES allows for three diverse key sizes of 128, 192, or 256 bits. The processing required 10 rounds when encrypting with the 128-bit key, 12 rounds for the 192-bit key, and 14 rounds for the 256-bit key. When ciphering with the 128 bit key, all 2^{128} keys mixtures must be inspected by decrypting the encrypted text with every one of those values [13].

For decryption and encryption, every round had four functions, excluding the final round with three. Encryption had the following functions: SubByte, ShiftRows, MixColumn, and AddRoundKey. A similar number of round functions were used for decryption, but with the opposite transformation.

The algorithm takes into account the diverse runtime periods with the same inputs, based on the speed of the processor, disk speed, instruction set, and type of compiler. The measured runtime "T (n)" is the amount of primary steps, taking into account that every progression step requires steady time. Since the inner loop iterates, the runtime was computed as:

$$T (n) = O (n^2) \qquad (1)$$

Where **n** is the size of the input data. Thus, the performance of an algorithm is proportional to the square of **n**. The study tested the performance impact of including encryption steps to the key-value database of web storage. Fig. 4 shows the results.

The results indicate that the process with encryption steps became slower by 10-40%. Therefore, applying encryption/decryption increased security but decreased response time. The process of encryption/decryption depended on hardware optimization and configuration. Devices with superior processors and greater memory storage sped up the process. AES functionality is now integrated into many processors, which helps to reduce encryption\decryption time.

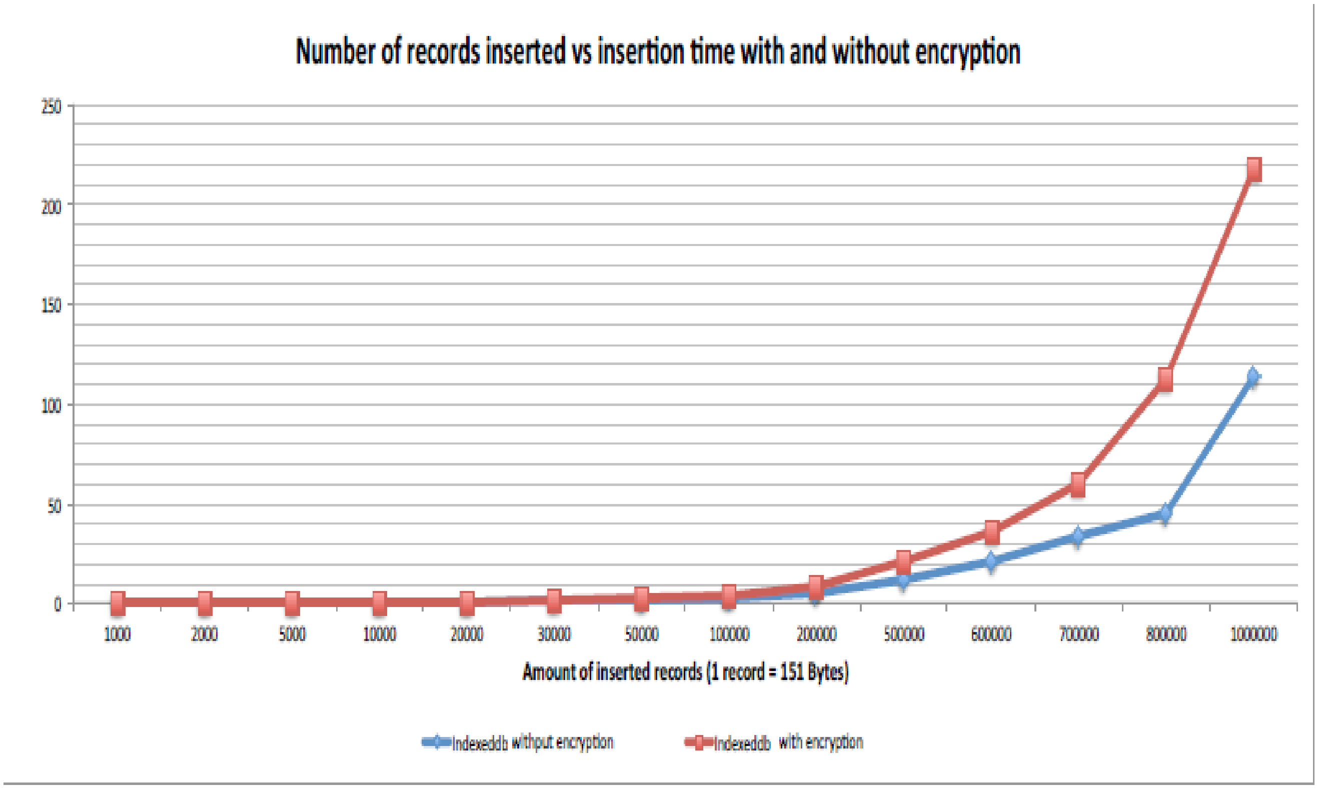


Fig. 4. The Time to Insert Data with and without Encryption into IndexedDB.

VIII. Conclusion and Future Work

Local browser storage has important advantages, when compared with a server-side database, including fast response, offline usage, and decreased network latency. However, currently, local browser storage is not secure. The main security concern is that data saved locally is unencrypted. The literature confirmed that the unencrypted data was not the only problem facing local browser storage. Data recovery of previously deleted data was another serious problem. Current security methods do not provide adequate security defenses for data saved locally, particularly data that may contain private information.

This study proposed and implemented a new security model for local browser storage. The new model added an additional layer amongst local browser storage and the web browser itself. The model contained an algorithmic framework for improving security against vulnerabilities identified in this study. JSCL was used in the proposed algorithm and was applied to the web browser as an extension. The SJCL library was chosen because SJCL is a secure, small, fast, multi-platform and powerful library for cryptography in JavaScript. The study combined the RSA standard with an SJCL library, which resulted in a hybrid encryption algorithm to guarantee data integrity.

When browser storage received a request from a web application to save data, the data was encrypted by the proposed browser extension. Encrypting data in browser storage kept data from being undermined regardless of whether the attackers acquired physical access to the machine, which may occur if a phone, tablet, or computer is lost or stolen.

Although the proposed model has been effectively applied to local browser storage, additional enhancements could be made in the future by extending the performance and security model. For example, a complete security library could be written to utilize diverse JavaScript crypto libraries. Further experiments with different crypto libraries could also be performed to compare the results.

Acknowledgments

We would like to express our deep appreciation to the people at Isra University who have helped us throughout our research.

References

[1] S.Sheikh, and M. AdeelPasha,"Energy-Efficient Multicore Scheduling for Hard Real-Time Systems: A Survey," ACM Transactions on Embedded Computing Systems (TECS), vol. 17, no. 6, 2018.

[2] N.Tahmasbi, and E. Rastegari, "A Socio-Contextual Approach in Automated Detection of Public Cyberbullying on Twitter," ACM Transactions on Social Computing, vol. 11, no. 4, 2018.

[3] P.Gradinger, D. Strohmeierand C. Christiane, "Definition and measurement of cyberbullying," Journal of Psychosocial Research on Cyberspace , vol. 4, no. 2, 2015.A. Robert, and J. Ravenscroft, "Dynamic data structures, a web based tool for teaching linked lists and binary trees," Journal of Computing Sciences in Colleges, vol. 33, no. 6, pp. 97-106, 2018.

[4] T. Al-Rousan, and H. Al Ese, "Impact of Cloud Computing on Educational Institutions: A Case Study," Recent Patents on Computer Science, vol. 8, no. 2, pp. 106-111, 2015.

[5] M. Díaz, M.Martín, and B.Rubio, "State-of-the-art, challenges, and open issues in the integration of Internet of things and Cloud Computing,"

[6] Journal of Network and Computer Applications, vol. 67, no. 1, pp. 99-117, 2016.

[7] V. Karavirta, and C. Shaffe, "Creating engaging online learning material with the JSAV JavaScript algorithm visualization," IEEE Transactions on Learning Technologies, vol. 9, no. 2, pp. 171--183, 2016.

[8] I. Koren, and R. Klamma, "The Exploitation of OpenAPI Documentation for the Generation of Web Frontends," in Proceedings of the The Web Conference 2018, Lyon, France, 2018.

[9] N. Correia, and J. Kleimola, "Web browser as platform for audiovisual performances," in Proceedings of the 11th Conference on Advances in Computer Entertainment Technology, Funchal, Portugal, 2017.

[10] A. McDonald, "Cookie confusion: do browser interfaces undermine understanding?," in Proceeding on Human Factors in Computing Systems, Atlanta, Georgia, USA, 2015.

[11] W.Kuang, L. Wu, and Y.Liu, "Key Selection for Multilevel Indices of Large-scale Service Repositories," in Proceedings of the 10th International Conference on Utility and Cloud Computing, Austin, Texas, USA, 2016,pp. 139-144.

[12] T. Al-Rousan, " An Investigation of User Privacy and Data Protection on User-Side Storage," International Journal of Online Engineering, 2019, vol. 15 no. 9, pp.17-30. 14, 2019.

[13] J.Hamilton, "Internet scale storage," in Proceedings of the 5th ACM SIGMOD International Conference on Management of data, Athens, Greece, 2018.

[14] M. Arenas, "Database Theory Column Report on PODS 2018," ACM SIGACT News, vol. 49, no. 4, pp. 55-57 , 2018.

[15] H. Oosterhuis, J. Culpepper, and M. Rijke, "The Potential of Learned Index Structures for Index Compression," in Proceedings of the 23rd Australasian Document Computing Symposium, Dunedin, New Zealand, 2018.

[16] R.Brunel, N. May, and A.Kemper, "Index-assisted hierarchical computations in main-memory RDBMS," VLDB Endowmen, vol. 9, no. 12, pp. 1065-1076 , 2016.

[17] L. Kim, and H. Lee , "Web-in-the-loop simulation framework for supporting CORS-based development," in Proceedings of the Poster Session and Student Colloquium Symposium, Alexandria, Virginia, 2017.

[18] J. Laassiri , "Data Security and risks for IoT in intercommunicating objects," in Proceedings of the 2nd international Conference on Big Data, Cloud and Applications, Tetouan, Morocco, 2017.

[19] J. Bozic, and F. Wotawa, "XSS pattern for attack modeling in testing," in Proceedings of the 8th International Workshop on Automation of Software Test, San Francisco, California, 2016.

[20] T. Al-Rousan, "Cloud Computing for Global Software Development: Opportunities and Challenges," in ransportation Systems and Engineering: Concepts, Methodologies, Tools, and Applications, IGI Global, 2015, pp. 897-908.

[21] N. Meng, S. Nagy, D. Yao, and D. Zhuang, "Secure coding practices in Java: challenges and vulnerabilitie," in Proceedings of the 40th International Conference on Software Engineering, Gothenburg, Sweden, 2018.

[22] H.Shahriar, "Security vulnerabilities and mitigation techniques of web applications," in Proceedings of the 6th International Conference on Security of Information and Networks, Aksaray, Turkey , 2016.

[23] W. Groef, F. Massacci, and T. Piessens, "NodeSentry: least-privilege library integration for server-side JavaScript," in Proceedings of the 30th Annual Computer Security Applications Conference, Louisiana, USA , 2017.

[24] M. Olan, "HTML5 jumpstart," Journal of Computing Sciences in Colleges, vol. 28, no. 3, pp. 35-36 , 2016.

[25] C. Aggarwal, Y. Li, P. Yu, and R. Jin, "On dense pattern mining in graph streams," VLDB Endowment, vol. 3, no. 1-2, pp. 975-984 , 2010.

[26] J. Oh, N. Son, and S. Lee, "A Study for Classification of Web Browser Log and Timeline Visualization," Lecture Notes in Computer Science , 2011.

[27] D. Ohana, and N. Shashidhar, "Do Private and Portable Web Browsers Leave Incriminating Evidence? A Forensic Analysis of Residual Artifacts from Private and Portable Web Browsing Sessions," Journal on Information Security, vol. 10, no. 1, pp. 135-142, 2013.

[28] H. Said, N. Al Mutawa, and I. Al Awadhi, “Forensic analysis of private browsing artifact, “Proceedings of the 11th Conference on Forensic analysis of privatbrowsing artifacts," 2011.

[29] K. Satvat, M. Forshaw, F.Hao, and E.Toreini, "On the privacy of private browsing –A forensic approach. ,," Journal of Information Security and Applications, vol. 19, no. 1, pp. 88-100, 2014.

[30] D. Ruiz, F. Amatte, and K. Park, "Overconfidence: Personal Behaviors Regarding Privacy that Allows the Leakage of Information in Private Browsing Mode," International Journal of Cyber-Security and Digital Forensics , vol. 4, no. 36, pp. 104-416, 2015.

[31] R. Montasari, and P. Peltola, "Computer Forensic Analysis of Private Browsing Modes," in Proceedings of the 10th Conference on Global Security, Safety and Sustainability: Tomorrow's Challenges of Cyber Security, Springer International Publishing, 2015.

[32] X. Gao, Y. Yang, H. Fu, and J. Lindqvist, "Private Browsing: an Inquiry on Usability and Privacy Protection," in Proceedings of the 15th Conference on Privacy in the Electronic Society, ACM.

[33] N.Tsalis , N. Virvilis, and A. Mylonas, "Browser Blacklists: A utopia of phishing protection. Security and Cryptography," in Security and Cryptography (Eds.), Lecture Notes (CCIS), Springer, 2017.

[34] S. Belloro, and A. Mylonas, "Security considerations around the usage of client-side storage APIs," Technical Report No. BUCSR-2018-01, 2018.

[35] K. Basques, "Tools for Web Developers - Performance Analysis Reference," Google, [Online]. Available: Performance Analysis Reference. [Online]. [Accessed 3-12-2018].

A Novel Approach for Dimensionality Reduction and Classification of Hyperspectral Images based on Normalized Synergy

Asma Elmaizi[1*], Hasna Nhaila[2], Elkebir Sarhrouni[3], Ahmed Hammouch[4], Nacir Chafik[5]
Research Laboratory in Electrical Engineering LRGE,
Mohammed V University, Rabat, Morocco

***Abstract*—During the last decade, hyperspectral images have attracted increasing interest from researchers worldwide. They provide more detailed information about an observed area and allow an accurate target detection and precise discrimination of objects compared to classical RGB and multispectral images. Despite the great potentialities of hyperspectral technology, the analysis and exploitation of the large volume data remain a challenging task. The existence of irrelevant redundant and noisy images decreases the classification accuracy. As a result, dimensionality reduction is a mandatory step in order to select a minimal and effective images subset. In this paper, a new filter approach normalized mutual synergy (NMS) is proposed in order to detect relevant bands that are complementary in the class prediction better than the original hyperspectral cube data. The algorithm consists of two steps: images selection through normalized synergy information and pixel classification. The proposed approach measures the discriminative power of the selected bands based on a combination of their maximal normalized synergic information, minimum redundancy and maximal mutual information with the ground truth. A comparative study using the support vector machine (SVM) and k-nearest neighbor (KNN) classifiers is conducted to evaluate the proposed approach compared to the state of art band selection methods. Experimental results on three benchmark hyperspectral images proposed by the NASA "Aviris Indiana Pine", "Salinas" and "Pavia University" demonstrated the robustness, effectiveness and the discriminative power of the proposed approach over the literature approaches.**

Keywords—Hyperspectral images; target detection; pixel classification; dimensionality reduction; band selection; information theory; mutual information; normalized synergy

I. Introduction

In the next decade, the exploitation of hyperspectral imaging [1] will experience a spectacular development thanks to the technological imaging evolution growing in many areas. The current generation of hyperspectral sensors provides large quantities of precise information on the nature and spatial-temporal evolution of the analyzed areas. The maturity and accessibility of this technology make it possible to address new applications in the fields of agronomy, environment, military, industrial and health security, etc. In remote sensing [2], the rich and detailed spectral information provided by hyperspectral images helped in detecting the composition of imaged materials and classifying targets with high spectral and spatial accuracy [3]. Embedded on an aircraft, a hyperspectral sensor operating in the visible near-infrared range (400-1000 nm) can simultaneously record several tens, even hundreds of narrow spectral bands. The volumes of data (data cubes) acquired often reach gigabytes for a single scene observed. As a result, their exploitation with classical methods developed for monochrome or color is very limited. In many cases, it is unnecessary to process all the spectral bands of an HSI [4][5] (Hughes phenomenon).

Most materials have specific characteristics only at certain bands, which makes the remaining spectral bands somewhat redundant. Additionally, some noisy bands [6] are influenced by various atmospheric effects. To overcome these challenges and respond quickly to the needs arising from the different potential applications, dimensionality reduction is an essential pre-processing step. Methods of bands selection must be developed to achieve the best compromise between reducing and preserving the amount of information acquired.

The selection approaches [7] consist of retaining the dataset physical meaning by selecting the most relevant bands. The hyperspectral band's selection will be the main topic of the work presented in this paper. Currently, selection algorithms can be categorized into two common approaches: wrapper and filters [8].

- The wrapper methods are classifier-dependent. They evaluate the band's relevance based on the classification accuracy and generally reach promising results. However, these approaches are very expensive in terms of computational complexity and may suffer from over-fitting to the learning algorithm.
- The filter methods are classifier-independent. They are based on the maximization of a certain evaluation function. The main advantages of these methods are their computational efficiency, simplicity and independence from the classifier. A common drawback shared by these approaches in literature is the lack of information about the synergy and interaction correlation between the picked bands and the ground truth.

In literature, many filter-selection methods have been developed using different evaluation measures. The evaluation function is generally based on distance, information, correlation and different consistency measures.

[*] Corresponding Authors

Information theory introduced by “Cover & Thomas” [9] has been widely applied in filter methods, where information measures are used to evaluate the band’s relevance and quantify the amount of information contained on images. This paper contributes to the knowledge in the area of hyperspectral dimensionality reduction by proposing a new approach based on normalized synergic correlation. The proposed method aims to overcome the limitations of the current state of the art filter band selection methods such as overestimation of the band significance, which causes selection of redundant and irrelevant bands. The new evaluation method selects the band that has maximum relevance, minimum redundancy and maximum normalized synergy with the previously selected bands. This paper reviews the state of art band selection methods highlighting their common limitations and comparing their performance versus the proposed algorithm. Experimental results are carried out using three benchmark hyperspectral images proposed by the NASA “AVIRIS Indiana Pine” [10], “Pavia University” and “ROSIS Salinas” [11]. Classification results are generated using the SVM [12][13] and KNN [14] to demonstrate the effectiveness and classification accuracy improvement of the proposed approach.

The rest of the paper is structured as follows. Section 2 describes the fundamentals of information theory and reviews the state of art band selection methods. Section 3 presents the proposed normalized max synergy (NMS) algorithm. Section 4 outlines the experiment conducted on the three datasets and analysis the achieved results. Finally, Section 5 concludes the paper.

II. Background on Information Theory based Approaches

In this section, we describe some basic concepts about information theory and feature selection, which will be used to build the proposed hyperspectral band selection algorithm.

The information theory proposed by "Cover & Thomas" [9] has been widely applied in filtering methods, where information measures are used to assess the relevance and discrimination of the characteristic.

Definition 1: The Shannon entropy introduced in (1) is defined as the quantification of the amount of information contained in variable X.

$$H(X) = \sum_{X} \mathrm{P(X)}\log_2 \mathrm{P(X)dX} \qquad (1)$$

Since, Shannon entropy H(X) is defined for a single variable and it is independent of the class, the mutual information between two random variables was introduced in order to measure the statistical dependence between the features and between the features and the class.

Definition 2: The mutual information (MI) of a pair of variables in (2) represents their degree of dependence in the probabilistic sense. It is the reduction of uncertainty on a random variable through the knowledge of another.

$$\mathrm{MI(X,Y)} = \sum_{\mathrm{X,Y}} \mathrm{P(X,Y)}\log_2 \frac{\mathrm{P(X,Y)}}{\mathrm{P(X)P(Y)}} \mathrm{dXdY} \qquad (2)$$

$$\mathrm{MI(X,Y\,)} = \mathrm{H(X)} + \mathrm{H(Y\,)} - \mathrm{H(X,Y\,)} \qquad (3)$$

$$\mathrm{MI(X,Y\,)} = \mathrm{H(X)} - \mathrm{H(X|Y)} \qquad (4)$$

The P(X,Y) in (2) is the joint probability function and P(X) , P(Y) represent the marginal probabilities.

In the equation (3), H(X) and H(Y) are the Shannon entropies of two variables X, Y respectively and H(X, Y) is the joint entropy between the variables. The mutual information can also be formulated using the conditional entropy as presented in (4).

Mutual information has the following properties.

- Mutual information is positive or zero.
- The mutual information is symmetrical.

In a wide survey of the feature selection literature, we have identified different information theory-based filters [15] and we will be presenting a selection of the most well-known criteria.

In the results section, the selected relevant methods will be applied to hyperspectral data to compare it with our proposed approach.

Battiti [16] proposed to use mutual information for variable selection in the Mutual Information-based Feature Selection (MIFS) algorithm. In this approach, the number of variables is fixed in advance and at each step, the variable that maximizes the mutual information between all the variables already selected is chosen. Formally, the variable selected by the MIFS algorithm is the one that maximizes the following goal function:

$$\mathrm{MIFS} = \mathrm{Argmax}(\mathrm{MI(Fi,C)} - \beta \sum_{\mathrm{Fs}\in\mathrm{S}} \mathrm{MI(Fi,Fs)}) \qquad (5)$$

The factor ‘β’ in (5) allows to control the redundancy term MI(Fi,Fs) and has a great influence on the selection algorithm. Several authors like Bollacker and Ghosh [17] that use different values for the parameter β without any justification. The value of β is often determined experimentally and depends on the data used. The problem is highlighted when the subset is very large and the redundancy term becomes larger than the relevance term. The algorithm will then select irrelevant features because they are not redundant, but not because they are relevant to the class.

As a consequence, several variants of the MIFS algorithm have been proposed in recent years in order to overcome its limitations. Kwak and Choi [18] proposed the algorithm MIFS-U as an improvement of MIFS.

$$\mathrm{MIFSU} = \mathrm{Argmax}(\mathrm{MI(Fi,C)} - \beta \sum_{\mathrm{Fs}\in\mathrm{S}} \frac{MI(C,Fs)}{H(Fs)} \mathrm{MI(Fi,Fs)} \quad (6)$$

Peng [19] analyzed as well the limitations of the previous selection approach and proposed a robust approach minimum redundancy maximum relevance (mRMR) where the redundancy term in (7) is divided over the cardinality of the subset.

$$\mathrm{mRMR} = \mathrm{Argmax}(\mathrm{MI(Fi,C)} - \frac{1}{\mathrm{S}} \sum_{\mathrm{Fs}\,\in\mathrm{S}} \mathrm{MI(Fi,Fs)}) \qquad (7)$$

Asma et al. [20] proposed a hybrid strategy combining the filter mRMR with the Fanno based wrapper strategy in order to select the relevant hyperspectral bands. Yang and Moody [21] proposed the Joint Mutual Information (JMI) (10) based on maximizing the cumulative summation of Joint Mutual Information of the selected subset.

The joint mutual information is presented in (9):

$$\mathrm{I(X,Y|Z)} = \ \mathrm{H(X|Z)} - \mathrm{H(X|Z,Y)} \qquad (8)$$

$$\mathrm{JMI(X,Y;Z)} = \ \mathrm{I(X,Z|Y)} + \mathrm{MI(X,Y)} \qquad (9)$$

The JMI filter approach is defined in (10) as follow:

$$\mathrm{JMI} = \mathrm{Argmax} \sum\nolimits_{\mathrm{Fs \in S}} \mathrm{JMI(Fi,Fs;C)} \qquad (10)$$

Meyer and al. introduced the Double Input Symmetrical Relevance (DISR) based on the joint mutual information as well [22]. The goal function of this approach is based on the symmetrical relevance as illustrated in (11).

$$\mathrm{DISR} = \mathrm{Argmax}\left(\sum\nolimits_{\mathrm{Bs \in S}} \frac{I(Fi,Fs;C)}{H(Fi,Fs;C)}\right) \qquad (11)$$

Within information theory studies dedicated to hyperspectral images selection, GUO [23] proposed an effective MI-based filter algorithm to select the most discriminative bands. He calculates the average of bands 170 to 210 of the HSI AVIRIS 92AV3C image that will be introduced in the result part in order to produce an estimated ground truth map (class), and use it instead of the real ground truth. Sarhrouni [24] proposed a mutual information based filter approach (MIBF) considering that the band maximizing the mutual information with the class or ground truth is a good approximation of it and introduced a threshold “Th” in order to control the redundancy criteria.

To select the most relevant bands, Nhaila presented also an enhanced version of the normalized mutual information [25]. This approach uses normalized MI to control the redundancy instead of MI as defined in the equation. This method is reported to perform well in terms of classification accuracy and stability.

III. Proposed Filter Approach based on Maximum Normalized Synergy MNS

A. Limitation of State of Art Methods

According to the previous part of this article, there are two factors that affect bands selection: MI between the bands and the ground truth (relevancy term) and MI between the selected bands (redundancy term). A remarkable weakness is that the majority of the feature selection algorithms did not consider the synergic dependence between the candidates bands with the other bands already selected. The methods discussed previously work on the assumption that the relevance of a single band is associated with the degree of dependence of this band on the ground truth. But it may happen that some HSI bands acting independently do not provide any additional information to the classification but when grouped together with other images gives promising results.

B. The Proposed Approach Normalized Mutual Synergy (NMS)

Aiming at the shortcomings of the above algorithms, a new band selection algorithm is proposed as an enhancement of the state of art methods. The proposed approach considers three factors: relevancy, redundancy and normalized synergy in order to select the discriminative bands. The purpose of the proposed approach is to reformulate the band selection problem as a modelling problem based on multi-criteria. We formalize this multi-criteria into three types of interaction between hyperspectral bands.

- Relevance criteria

The decision of which bands are the most relevant and should be selected is usually associated with the degree of dependency of each single band to the ground truth (class). This criterion is calculated using the mutual information between the selected band and the ground truth MI(Bi,GT) and it is reflecting the shared information between the selected band Bi and the ground truth and will be used to evaluate the discriminative ability of each band to the classification.

- Redundancy criteria

This criterion is a reflection of the common information shared by the selected bands. The amount of redundancy can never decrease when other new bands are added. The redundancy will be controlled using the normalized mutual synergy NMS that will be presented in the next part.

- Complementary criteria

The decision of which bands are the most discriminative and powerful for target classification is usually associated with the degree of complementarity and synergy between the selected bands. The complementarity will be evaluated using the normalized interaction information NMS. This measure will be used to control simultaneously the redundancy and the complementarity between the picked bands.

In fact, Guo [23] and Sarhouni [24] uses MI(Fi,Fs) to measure the bands redundancy. They considered that all correlated bands are redundant but neglect that some of the wrongly removed bands are synergic. The normalized mutual information algorithm (NMI) [25] was proposed as an enhancement of mutual info based methods. According to the NMI algorithm, there are two kinds of bands correlations: Independent correlation when the measure is 0 and redundant correlation when the measure is between 0 and 1. This approach considers all the correlated bands as redundant and would wrongly judge the synergic bands as redundant as well. To overcome the limitations discussed previously, we propose the NMS algorithm that can provide a more accurate measure for band interaction including the three criteria (relevance, redundancy and complementarity).

Definition 3: The interaction information I(X;Y;Z) or synergy S(X,Y) has been defined by Jakulin [26] as the decrease in uncertainty caused by joining the attributes X and Y in a Cartesian product [27]. Considering the bands and the class label simultaneously, the bands synergy in (12) can be defined as follows:

$$\mathrm{S(X,Y)} = \mathrm{I(X,Y,Z)} = \mathrm{I(X,Y)} - \mathrm{I(X,Y|Z)} \qquad (12)$$

Substituting eq. (3) for I(X,Y) and I((X,Y)|Z) in (13)

$$(\mathrm{I(X,Y,Z)} = [\mathrm{H(X)} + \mathrm{H(Y)} - \mathrm{H(X,Y)}] - [\mathrm{H(X|Z)} + \mathrm{H(Y|Z)} \ - \mathrm{H((X,Y)|Z]} \qquad (13)$$

From (13) we can deduce the relationship between mutual information and the synergy measure as below:

$$\mathrm{S}(\mathrm{X},\mathrm{Y}) = \mathrm{I}(\mathrm{X},\mathrm{Y},\mathrm{Z}) = \mathrm{I}\big((\mathrm{X},\mathrm{Y}),\mathrm{Z}\big) - \mathrm{I}(\mathrm{X},\mathrm{Z}) - \mathrm{I}(\mathrm{Y},\mathrm{Z}) \qquad (14)$$

The normalized mutual synergy is defined in (15) as follow:

$$\mathrm{NMS}(\mathrm{X},\mathrm{Y},\mathrm{Z}) = \frac{2\,\mathrm{I}(\mathrm{X},\mathrm{Y},\mathrm{Z})}{\mathrm{I}(\mathrm{X},\mathrm{Z})+\mathrm{I}(\mathrm{Y},\mathrm{Z})} \qquad (15)$$

Proposed Algorithm Normalized mutual synergy NMS

Inputs :

- Hyperspectral Dataset H
- S={b1,b2,...,bn} Spectral bands of the HSI image
- GT: the class label or ground truth that will be used for supervised classification
- A defined percentage of pixels for training and pixels for testing.

Outputs:

- The subset R of k selected bands ranked in the selection order.
- Reproduced ground truth.
- Classification results and metrics.

1. Set R 🡐 [ ] "empty result set in the initialization step"
2. Select the relevant band that got the max mutual information with the class label (ground truth)

   b*=Argmax bi€S (MI (bi,GT))

   Set S🡐S-{b*], R🡐{b*}

3. Calculate the ground truth estimated using the first selected band

   Set GTest=b*

4. While |R|≤ k during each iteration choose the band that maximizes the objective function:

   $$\textbf{F(b*)= Argmax}\big(\mathbf{MI}(\mathbf{bi},\mathbf{GT}) + \mathbf{NMS}(\mathbf{bi},\mathbf{GTest},\mathbf{GT})\big)$$

   $$\textbf{F(b*)= Argmax}(\mathbf{MI}(\mathbf{bi},\mathbf{GT}) + \mathbf{2} * \frac{\mathbf{I(bi,GTest,GT)}}{\mathbf{MI(bi,GT)+MI(GTest,GT)}})$$

5. Recalculate the class label estimated using the new selected band in each iteration:

   $$\mathbf{GTest} = \frac{\mathbf{GTest + b*}}{\mathbf{2}}$$

   S🡐S-{b*], R🡐RU{b*}

6. Output the result subset R containing the selected bands.
7. Evaluate the selected bands using a classifier function.
8. Generate the classification metrics and the reproduced ground truth.

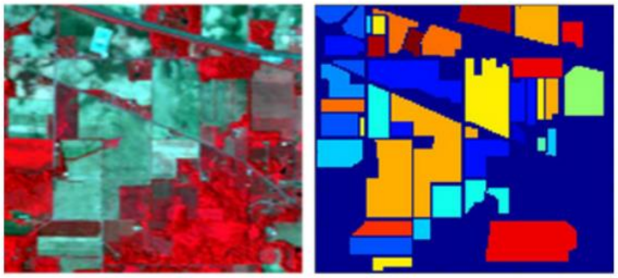


Fig. 1. Proposed Normalized Synergy Algorithm (NMS).

In the proposed NMS filter approach, the normalized mutual synergy criteria will be used to evaluate the redundancy and complementarity between bands as follow:

- The NMS measure is positive: 0<NMS≤1 when the selected bands have synergic correlation and together provide critical info for accurate classification which cannot be provided by each one of them individually.
- The NMS measure is negative: -1≤NMS<0 when the selected bands are redundantly correlated and provide redundant information that does not help to increase the classification accuracy.
- The NMS measure is equal to zero when the selected band is independent from the already selected bands in the context of the class label of ground truth.

Let S={b1,b2,...,bn} the band's set of hyperspectral image dataset H with n bands. The goal is to find a subset of bands that maximizes the objective function f(x) based on the NMS measure and which represents the combination of these three types of interaction. In order to avoid testing all possible combinations that will cause a computational burden, we propose a greedy selection algorithm (Fig. 1) that begins with an empty set of bands. The first chosen band will be based on the relevance criteria and thus will be the one with the maximum mutual information with the ground truth and successively adds bands during each iteration that maximize the objective function f(x) combining the three criteria. Afterward, the selected band's classification rate will be evaluated using the classifiers Support Vector Machine (SVM) and K-Nearest (KNN).

IV. Experimental Results and Analysis

In order to evaluate the performance of the proposed filter method, we will use three real hyperspectral datasets from NASA's Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS) [10] and the Reflective Optics System Imaging Spectrometer (ROSIS) [11]. These images are one of the most challenging classification problems since they are overlaid with mixed pixels and similar classes and they will be presented shortly on this section. Then, we will introduce as well the classifiers and evaluation metrics that will be used for results analysis. Finally, the experimental results for each hyperspectral image are presented and discussed by comparing with the close literature methods.

A. *Experimental Datasets*

1) Aviris indiana pines: The Indiana pines [10] is an agricultural image acquired over the Indian Pines test site in North-western Indiana, USA and collected by the Airborne Visible Infra-Red Imaging Spectrometer. The Aviris Indiana pines hyperspectral image is built using a 3-dimensional cube with two spatial and a third spectral as illustrated in Fig. 2.

Every material and compound of the earth surface illustrated in the ground truth (Fig. 3) is identified with its unique electromagnetic signature. It consists of 224 spectral reflectance bands in the wavelength range 0.4–2.45 µm covering 145*145 pixels. This scene contains two-thirds agriculture (alfalfa, corn, oats, soybean, wheat), and one-third forest (woods, and different sub-classes of grass or other natural perennial vegetation). The ground-reference data of the scene is designated into 16 classes with a total of 10,366 labelled samples (Fig. 3).

2) Rosis pavia university: The Pavia scene was recorded using ROSIS (Reflective Optics System Imaging Spectrometer) [11] over the Pavia University in Italy. It consists of 103 spectral reflectance bands in the wavelength range 0.43- 0.86 µm covering 610*340 pixels. This scene covers an urban environment that is mainly constituted of Natural objects (trees, meadows, soil), various solid structures (asphalt, gravel, metal sheets, bricks) and shadows. The ground-reference data of the scene is designated into 9 classes as illustrated in Fig. 4.

3) Aviris salinas: The Salinas scene was acquired over the Salinas Valley in California and gathered by the AVIRIS sensor [10] as well. The scene consists of 224 spectral reflectance bands covering 512 × 217 pixels and it mainly contains covering vegetables, fields and bare soils. The ground-reference image of the scene is designated into 16 classes and presented in Fig. 5.

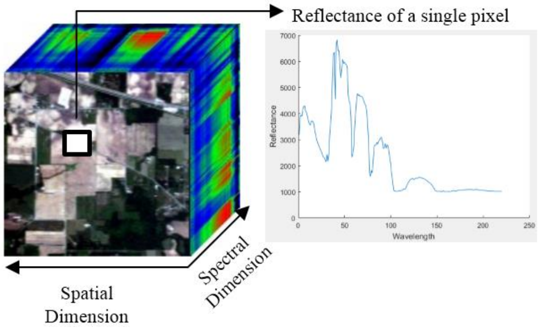


Fig. 2. Data Cube of the Hyperspectral, Image Aviris Indiana Pines.

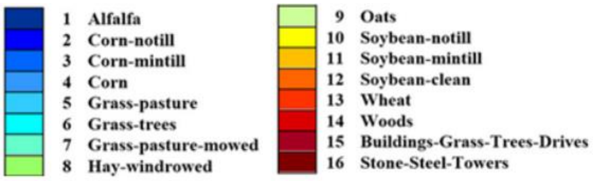


Fig. 3. (Right) Ground Truth Data of the Indian Pines image. (Left) Three-Band Color Composite of the Indian Pines Image (Bands 30, 43, and 21).

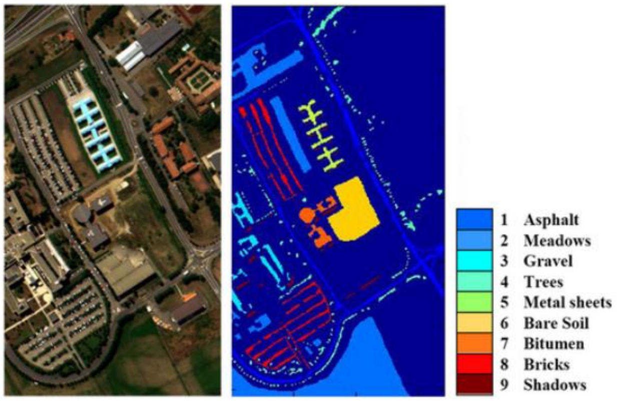


Fig. 4. (Right) Ground Truth Data of the Pavia University Image. (Left) Three-Band Color Composite of the Indian Pines Image (Bands 13, 33, and 56).

B. *Classifiers and Evaluation Metrics*

To assess the performance of the proposed method, the support vector machine (SVM) [12][13] and the k-nearest neighbor (KNN) [14] were chosen for the classification step. The Gaussian radial basis function (RBF) kernel is adopted for the SVM classifiers and the cross-validation operation is processed in order to determine the optimal parameters C and γ of the RBF kernel. The KNN algorithm is used with the Euclidean distance and k=3 nearest neighbors.

The SVM and KNN algorithms choice is based on our previous comparison study [28] where both classifiers results showed their great performance for HSI classification compared to other classifiers.

In all experiments, 10%, 25% and 50% of instances in each class are randomly labelled to compose the training sets and the remaining pixels are considered for the test and validation.

In order to evaluate the performance of the proposed method and compare it with other literature approaches, we calculate several well-known metrics in the literature reflecting the classification accuracy performance (OA, AA and KC).

- The Overall accuracy (OA) refers to the number of correctly classified instances divided by the total number of testing samples.
- The Average accuracy (AA) is a measure of the mean value of the classification accuracies of all classes.
- The kappa coefficient (KC) is a statistical measurement of consistency between the ground truth map and the final classification map.

C. *Classification Results and Discussion*

1) Classification results on HS image aviris indiana pine: Table I presents a comparison of classification results between the proposed approach (NMS) versus the information theory-based filters (MIBF, JMI, DISR and NMI). We carry out a set of parallel experiments in order to calculate the overall accuracy, the average accuracy and the kappa coefficient for the selected bands. Each column of Table I represents the 16 individual class accuracies of the Indiana scene using the

SVM and the KNN classifiers. The results reflect the robustness and strength of the proposed approach in selecting highly discriminative bands. In all cases, classification accuracies decreased when using the KNN instead of the SVM for the classification stage.

This result confirms the fact that the classifier SVMs are less affected by the Hughes phenomenon especially when trained with mixed spectral-spatial data confirming the results obtained on our previous work [28]. Fig. 6 presents the classification accuracy rate of the proposed approach with regard to the other band selection algorithms for different selected bands number up to 80 selected bands. From the Indiana pines results, we remarked that the MIBF algorithm has the lowest classification accuracy rate due to the weak band's correlation estimation. The redundancy term in this method is affected by the choice of the threshold Th and is often determined experimentally based on the dataset used. Additionally, this algorithm performance decreases when the subset is very large and the redundancy term becomes larger than the relevance term. The JMI and DISR algorithms provide promising results compared to MIBF and NMI due to the joint mutual info based objective function that gives better bands estimation.

The JMI reaches 91.49 % for 60 selected bands which is more than NMI by 2.86% and MIBF 2.32%. Our proposed approach outperform the other methods with (OA=94,09% & Kappa =93.69%, AA=94.3%) for 40 selected bands.

Experimental results reflected the effect of the normalized synergy adopted in the objective function of our algorithm. In fact, the correlation between bands is evaluated and classified into three types of interaction: relevancy, redundancy and synergic. It is worth noting also that the NMS proposed algorithm selects the high discriminative bands with a high speed as shown in Fig. 6 for small selected bands number.

During each iteration, the band is only selected when it increases the objective function based on the three correlation types.

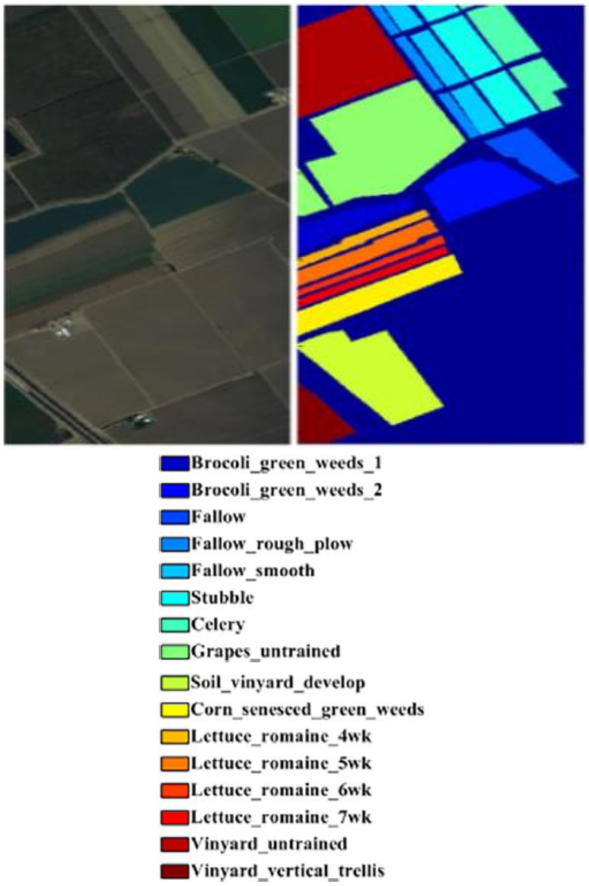


Fig. 5. (Right) Ground Truth Data of Salinas Image. (Left) Three-Band Color Composite of the Salinas Image (Bands 10, 30, and 21).

TABLE. I. CLASSIFICATION RESULTS OF THE PROPOSED APPROACH VERSUS DIFFERENT LITERATURE APPROACHES IN INDIANA PINES DATASET AND USING KNN & SVM (40 SELECTED BANDS)

Algorithm Indiana	**MIBF**		**NMI**		**JMI**		**DISR**		**NMS (Proposed approach)**	
Classifier	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**
1	68,52	81,48	55,56	74,07	64,81	87,04	66,67	87,04	83,33	**90,74**
2	76,43	75,24	73,85	70,08	74,55	71,2	76,78	74,06	77,75	**91,63**
3	78,9	80,22	72,3	71,82	66,31	66,19	70,5	78,18	76,14	**89,93**
4	61,54	75,64	60,26	76,92	58,12	77,78	66,24	84,19	73,5	**89,32**
5	84,91	89,13	83,1	83,5	94,16	97,38	94,57	98,19	94,57	**98,79**
6	95,05	94,91	95,18	92,24	98,39	98,39	98,13	98,53	97,46	**98,53**
7	42,31	76,92	26,92	69,23	50	84,62	65,38	80,77	80,77	**92,31**
8	96,32	97,75	95,71	95,91	96,93	98,57	97,55	98,36	98,57	**99,18**
9	55	95	85	90	80	90	80	90	80	**100**
10	70,97	72,52	73,97	74,07	67,15	54,55	70,56	64,67	88,12	**92,46**
11	83,71	85,74	83,1	85,25	84,48	86,39	85,49	87,52	88,7	**93,64**
12	76,06	88,27	71,5	82,9	68,57	74,59	70,52	75,57	73,62	**92,35**
13	97,17	98,11	95,28	95,75	99,06	99,06	98,58	99,53	98,11	**98,11**
14	94,44	94,36	91,73	94,82	96,14	98,15	96,75	98,3	96,45	**98,38**
15	52,89	59,21	51,58	54,74	62,37	81,84	63,16	81,58	57,89	**86,58**
16	86,32	96,84	82,11	96,84	84,21	94,74	86,32	95,79	93,68	**96,84**
Kappa	**80,71**	**83,31**	**78,93**	**80,73**	**80,05**	**81,23**	**81,87**	**84,26**	**85,38**	**93,69**
AA	**76,28**	**85,08**	**74,82**	**81,76**	**77,83**	**85,03**	**80,45**	**87,02**	**84,92**	**94,3**
OA	**81,92**	**84,35**	**80,25**	**81,93**	**81,3**	**82,4**	**83**	**85,24**	**86,29**	**94,09**

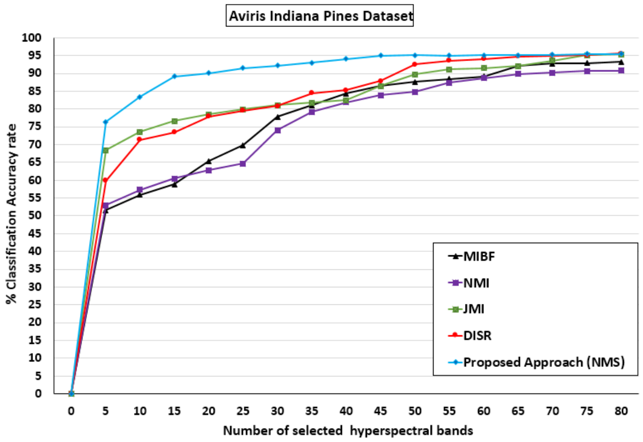


Fig. 6. The Classification Accuracy Rate Results Versus the Number of Selected Bands using the SVM Classifier for the Indiana Pines Image.

Fig. 7 presents reproduced classification maps using 40 selected bands using the proposed algorithm (NMS). The obtained result confirms that the selected 40 bands are highly discriminative to distinguish and classify the scene materials and target.

The classification accuracy reaches 94.09 % using the support vector machine classifier and we notice that 40 bands selected using the NMS approach are sufficient to detect all scene materials and provide a produced maps close to the ground truth.

2) Classification results on HS image rosis pavia: Table II illustrates the classification accuracy rate of the proposed algorithm compared to four literature approaches using SVM and KNN classifiers. Each row of the table provide the individual accuracies of the Pavia scene and the last three rows generate the overall, average and kappa classification metrics. From the produced result, we confirm the robustness of the proposed approach that outperforms the other mutual information-based filters for the Pavia scene. In fact, the normalized synergy method selects relevant bands rapidly due to the accurate objective function used during the selection process for both classifiers SVM and KNN. Our proposed method achieves an overall accuracy of 94.7% classification accuracy for 40 selected bands, which is higher than the JMI by 5.04% and DISR by 4.03. Fig. 8 present the evaluation of the proposed approach with regard to other features selection algorithm defined in the literature for the Pavia scene. The analysis of the results of the different curves confirms that the evaluation of bands correlation using the NMS helps in improving the results significantly compared with other algorithms. Fig. 9 illustrates the reproduced maps using 40 selected bands by the NMS. The dimensionality reduction of the pavia scene into 40 pertinent bands allows classification of image pixels and detection of material and target of the scene with high accuracy. The selected bands are enough to discriminate the material in 9 classes of the image and reproduce a close map to the ground truth.

3) Classification results on HS image aviris salina: The following Table III and Fig. 10 confirms the robustness of our proposed method that outperforms the other filters for this dataset as well. It is remarkable that the JMI and DISR algorithms in a second-place perform better than the other filters based on mutual information. This result is due to the joint mutual information objective function that provides an accurate bands evaluation. Using the KNN classifier, we achieved OA=92.56%, kappa=92.06% and AA=96.27% for 40 bands.

TABLE. II. CLASSIFICATION RESULTS OF THE PROPOSED APPROACH VERSUS DIFFERENT LITERATURE APPROACHES IN PAVIA UNIVERSITY DATASET AND USING KNN & SVM (40 SELECTED BANDS)

Algorithm	**MIBF**		**NMI**		**JMI**		**DISR**		**NMS (Proposed approach)**	
Classifier	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**
1	89,02	91,86	91,83	93,08	91,92	93,68	92,02	94,33	92,99	**94,06**
2	97,66	97,69	96,49	96,9	96,66	97,06	96,95	97,11	98,34	**98,34**
3	71,18	67,98	75,75	76,08	75,66	69,22	76,04	70,75	77,61	**77,7**
4	86,06	95,59	89,36	92,72	89,3	90,7	89,52	90,86	90,57	**96,44**
5	99,63	100	99,48	100	99,48	99,85	99,48	99,93	99,63	**100**
6	78,72	83,4	68,7	71,76	67,05	62	68,28	67,77	84,39	**92,24**
7	86,84	74,66	86,54	84,66	86,02	82,33	86,84	85,34	90,68	**81,88**
8	86,58	86,12	89,76	89,63	87,02	89,87	86,77	90,11	90,33	**90,36**
9	100	100	99,89	99,89	99,68	99,68	99,68	99,68	99,89	**99,79**
Kappa	**89,64**	**90,9**	**89,04**	**90,09**	**88,61**	**88,37**	**88,99**	**89,5**	**92,63**	**94,04**
AA	**88,41**	**88,59**	**88,65**	**89,41**	**88,09**	**87,16**	**88,4**	**88,43**	**91,6**	**92,31**
OA	**90,79**	**91,91**	**90,25**	**91,19**	**89,88**	**89,66**	**90,21**	**90,67**	**93,44**	**94,7**

TABLE. III. CLASSIFICATION RESULTS OF THE PROPOSED APPROACH VERSUS DIFFERENT LITERATURE APPROACHES IN SALINAS DATASET AND USING KNN AND SVM (40 SELECTED BANDS)

Algorithm Salinas	**MIBF**		**NMI**		**JMI**		**DISR**		**NMS (Proposed approach)**	
Classifier	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**
1	95,17	94,47	98,86	96,62	98,46	97,46	99,9	98,56	**99,35**	99,25
2	87,47	92,3	99,84	99,89	98,15	97,56	99,97	99,97	**99,89**	100
3	99,7	99,7	98,53	97,27	98,84	98,73	99,19	98,48	**99,49**	99,24
4	98,92	96,34	99,71	99,64	99,57	99,5	99,71	99,5	**99,64**	99,5
5	98,43	98,39	99,1	97,61	97,91	96,71	99,03	96,79	**99,33**	98,81
6	99,55	99,07	99,97	99,97	99,55	99,12	99,95	99,97	**99,95**	99,95
7	87,96	88,13	99,86	99,72	99,53	99,75	99,86	99,75	**99,44**	99,8
8	83,46	88,87	83,67	89,1	87,65	89,73	84,09	91,89	**85,25**	88,97
9	99,27	99,27	99,65	99,27	99,53	99,44	99,74	99,24	**99,56**	99,32
10	94,02	91,21	95,55	90,82	95,61	89,99	95,33	91,58	**95,76**	90,58
11	97,28	95,04	96,82	90,17	94,01	89,42	97,1	91,57	**98,31**	94,3
12	99,12	98,91	99,27	98,6	99,48	98,34	99,69	99,07	**99,79**	96,63
13	97,82	98,47	99,45	99,02	98,58	98,58	99,24	99,13	**98,69**	99,69
14	94,58	91,4	97,2	93,93	96,45	95,14	97,94	93,46	**95,14**	98,03
15	66,24	45,2	70,07	41,99	79,8	51,51	70,45	37,42	**72,36**	93,36
16	95,57	94,85	98,89	94,58	98,84	98,78	99,39	96,79	**98,28**	53,74
Kappa	**88,49**	**86,59**	**91,37**	**87,59**	**93,3**	**89**	**91,64**	**87,8**	**92,06**	**98,23**
AA	**93,41**	**91,98**	**96,03**	**93,01**	**96,37**	**93,74**	**96,29**	**93,32**	**96,27**	**94,93**
OA	**89,21**	**87,43**	**91,91**	**88,36**	**93,72**	**89,69**	**92,16**	**88,56**	**92,56**	**90,62**

Fig. 11 illustrates the reproduced ground truth for 40 bands using the SVM Classifier. Using the bands selected by our approach, we were able to detect the 16 classes included in the Salinas scene with an accuracy equal to 90.62.

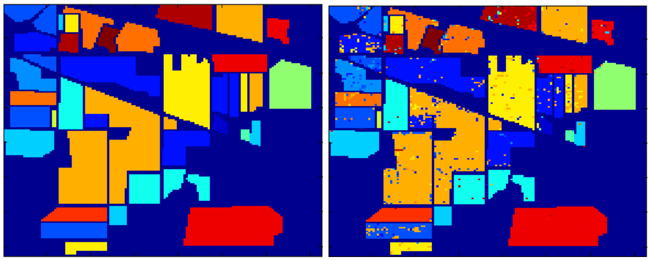


Fig. 7. Indiana Pines Ground Truth (Left), Indiana Ground truth Reproduced Map (Right) using the Proposed Algorithm NMS for 40 Bands (94.09%).

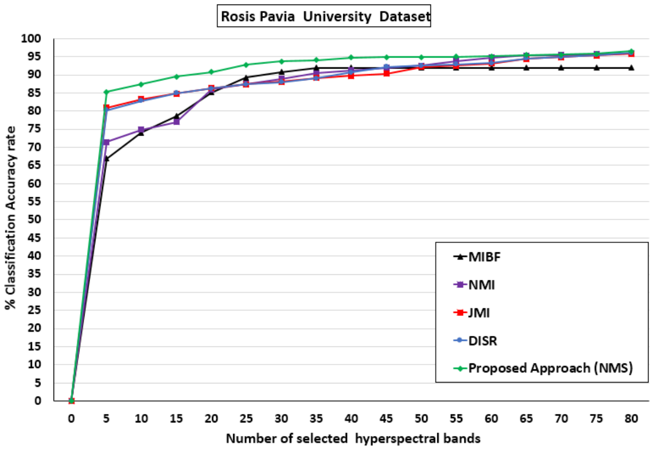


Fig. 8. The Classification Accuracy Rate Results Versus the Number of Selected Bands using the SVM Classifier for the Pavia University Image.

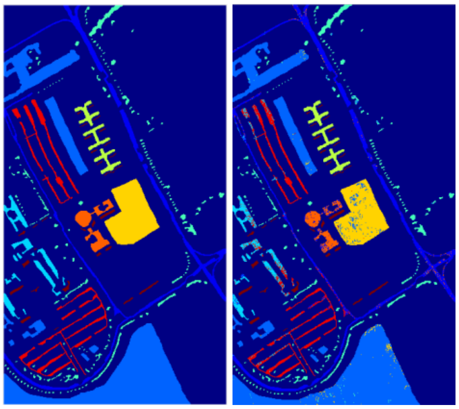


Fig. 9. University Pavia Ground Truth (Left), Pavia Reproduced Map (Right) using the Proposed Algorithm NMS for 40 Bands (OA=94.7%).

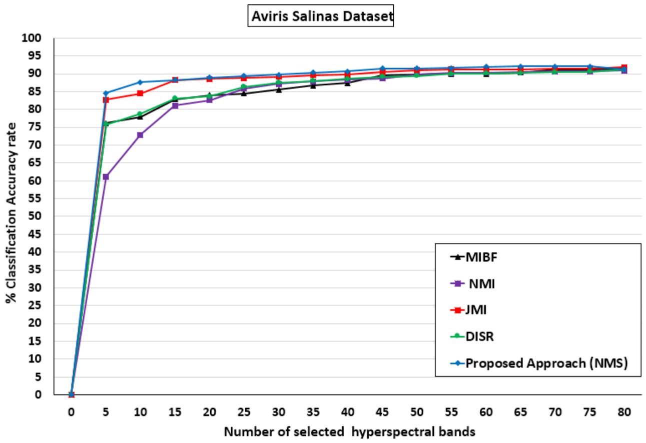


Fig. 10. The Classification Accuracy Rate Results Versus the Number of Selected Bands using the SVM Classifier for the Salinas Image.

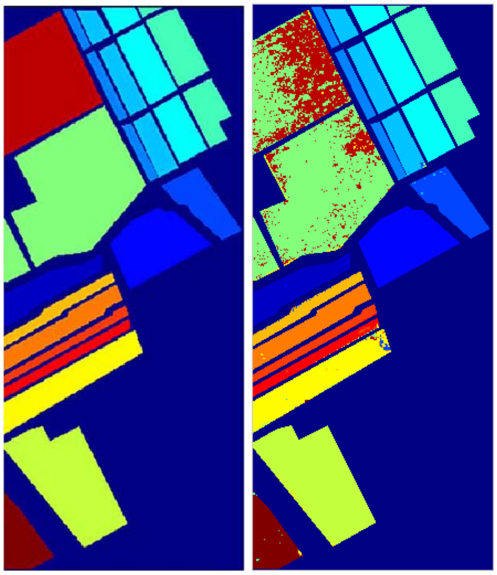


Fig. 11. Salinas Ground Truth (Left), Salinas Reproduced Map (Right) using the Proposed Algorithm NMS for 40 Bands (OA=90.62%).

V. Conclusion

In the last decades, remote sensing community has achieved a great improvement in detecting targets and classifying materials of difficult scenes due to the hyperspectral technology. However, the high dimensionality reduction of this type of images had always been a necessity in order to detect materials with high classification accuracy. In this paper, we present a new band selection approach based on information theory: normalized mutual synergy. This method is designed in order to resolve the problem of the high dimensionality of the hyperspectral images by selecting the discriminative bands, removing redundant and noisy ones. This method is based on the evaluation of every single band of the hyperspectral cube based on an objective function maximization. The evaluation function is a combination between the three correlation types: normalized synergy, redundancy and relevancy.

The robustness and effectiveness of the proposed approach have been evaluated using three hyperspectral public datasets by the NASA. Experimental results using the SVM and KNN classifiers confirm that the proposed approach increases the classification accuracy significantly and helps in selecting high discriminative bands rapidly. Compared to the other filter methods, our algorithm evaluates the band’s correlation and interaction with high accuracy in order to select the discriminative bands and helps in detecting the scene materials to provide a reproduced map close to the ground truth.

Future work includes more experiments using other hyperspectral datasets and including new spectral parameters in order to improve bands evaluation and significance.

References

[1] Khan, Muhammad Jaleed, et al. "Modern trends in hyperspectral image analysis: a review." IEEE Access 6 (2018): 14118-14129.

[2] Pham, T. T., et al. "Airborne Object Detection Using Hyperspectral Imaging: Deep Learning Review." International Conference on Computational Science and Its Applications. Springer, Cham, 2019.

[3] He, Lin, et al. "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines." IEEE Transactions on Geoscience and Remote Sensing 56.3 (2017): 1579-1597.

[4] Hughes, Gordon. "On the mean accuracy of statistical pattern recognizers." IEEE transactions on information theory 14.1 (1968): 55-63.

[5] Bellman, Richard E. Adaptive control processes: a guided tour. Vol. 2045. Princeton university press, 2015.

[6] Duan, Puhong, et al. "Noise-Robust Hyperspectral Image Classification via Multi-Scale Total Variation." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2019).

[7] Sarhrouni, Elkebir, Ahmed Hammouch, and Driss Aboutajdine. "A dashboard to analysis and synthesis of dimensionality reduction methods in remote sensing." Int. J. Eng. Technol 5.3 (2013): 2678-2684.

[8] Elmaizi, Asma, et al. "Hybridization of filter and wrapper approaches for the dimensionality reduction and classification of hyperspectral images." 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE, 2017.

[9] Cover, Thomas M. "Thomas. Elements of information theory." Wiley Series in Telecommunications (1991).

[10] Green, Robert O., et al. "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)." Remote sensing of environment 65.3 (1998): 227-248.

[11] Holzwarth, S., et al. "HySens-DAIS 7915/ROSIS imaging spectrometers at DLR." Proceedings of the 3rd EARSeL Workshop on Imaging Spectroscopy. 2003.

[12] Melgani, Farid, and Lorenzo Bruzzone. "Classification of hyperspectral remote sensing images with support vector machines." IEEE Transactions on geoscience and remote sensing 42.8 (2004): 1778-1790.

[13] Pal, Mahesh, and P. M. Mather. "Support vector machines for classification in remote sensing." International Journal of Remote Sensing 26.5 (2005): 1007-1011.

[14] Guo, Yanhui, et al. "K-Nearest Neighbor combined with guided filter for hyperspectral image classification." Procedia Computer Science 129 (2018): 159-165.

[15] Chandrashekar, Girish, and Ferat Sahin. "A survey on feature selection methods." Computers & Electrical Engineering 40.1 (2014): 16-28.

[16] Battiti, Roberto. "Using mutual information for selecting features in supervised neural net learning." IEEE Transactions on neural networks, 1994, 5.4: 537-550.

[17] Bollacker, Kurt D., and Joydeep Ghosh. "Mutual information feature extractors for neural classifiers." Proceedings of International Conference on Neural Networks (ICNN'96). Vol. 3. IEEE, 1996.

[18] Kwak, Nojun, and Chong-Ho Choi. "Input feature selection by mutual information based on Parzen window." IEEE Transactions on Pattern Analysis & Machine Intelligence 12 (2002): 1667-1671.

[19] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." IEEE Transactions on pattern analysis and machine intelligence 2005, 27.8:1226-1238.

[20] Elmaizi, Asma, et al. "Hybridization of filter and wrapper approaches for the dimensionality reduction and classification of hyperspectral images." 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE, 2017.

[21] Yang, H., and John Moody. "Feature selection based on joint mutual information." Proceedings of international ICSC symposium on advances in intelligent data analysis. 1999.

[22] Meyer, Patrick Emmanuel, Colas Schretter, and Gianluca Bontempi. "Information-theoretic feature selection in microarray data using variable complementarity." IEEE Journal of Selected Topics in Signal Processing 2.3 (2008): 261-274.

[23] Guo, Baofeng, et al. "Band selection for hyperspectral image classification using mutual information." IEEE Geoscience and Remote Sensing Letters, 2006, 3.4:522-526.

[24] Sarhrouni, ELkebir, Ahmed Hammouch, and Driss Aboutajdine. "Dimensionality reduction and classification feature using mutual information applied to hyperspectral images: a filter strategy based algorithm." arXiv preprint arXiv:2012, 1210.0052.

[25] Nhaila, Hasna, et al. "A Novel Filter Approach for Band Selection and Classification of Hyperspectral Remotely Sensed Images Using Normalized Mutual Information and Support Vector Machines." International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning. Springer, Cham, 2018.

[26] Jakulin, Aleks. Machine learning based on attribute interactions. Diss. Univerza v Ljubljani,2005.

[27] Jakulin, Aleks. Attribute interactions in machine learning. Diss. Univerza v Ljubljani ,2003.

[28] Nhaila, Hasna, et al. "Supervised classification methods applied to airborne hyperspectral images: comparative study using mutual information." Procedia computer science 148 (2019): 97-106.

Muscle Electro Stimulator for the Reduction of Stretch Marks

Paima-Sahuma Jaime[1], Duran-Berrio Linder[2], Palacios-Cusiyunca Chrisstopher[3], Roman-Gonzalez Avid[4]
Electronics and Telecommunications Department, Universidad Nacional Tecnológica de Lima Sur, Lima, Peru[1, 2, 3]
Aerospace Sciences and Health Research Laboratory (INCA-Lab), Universidad Nacional Tecnológica de Lima Sur, Lima, Peru[4]

***Abstract*—The problem of stretch marks is generated because the skin stretches abruptly in a short time; this change causes the skin to deform and widen, forming a roughness. This roughness is what is known as stretch marks. This document arose from the need to reduce the deformation in the skin that many people suffer. This deformation mainly is due to overweight, pregnancy, or during adolescence due to rapid growth. In this document, a device will be designed that will have the task of reducing skin roughness. One will use electro-stimulation as the primary technique to apply electrical impulses. This device can limit and control the electrical signals produced to control the movement of muscle fibers and skin. The results obtained show a remarkable reduction of stretch marks in one people after the application of electrical stimuli with the device. Research shows promising results.**

***Keywords*—Electro stimulator; stretch marks; muscle fibers**

I. Introduction

From a physiological point of view, stretch marks are caused by stretching of the middle layers of skin, internal skin due to pregnancy, puberty, rapid growth, and sudden weight gain or weight training.

Stretch marks are skin atrophies, very visible thinning of the skin. These marks occur as a result of breakage and partial loss of collagen fibers and elastin in the affected region. This situation resulting in decreased skin cohesion and yielding said area tension forces of muscle mass that support the skin [1].

Stretch marks are found mainly in the abdomen, legs, and breasts. In reality, they are a pathological state of the connective tissue of the dermis, isolated by excessive localized fibrosis in the form of ropes in response to breakage and poor fiber quality. It can be said that they are scars of this tissue. If stretch marks, when formed, are violet or purple, it indicates that the dermis still has a blood supply. If then they are roses, it also has irrigation.

When there is a breakdown of collagen and elastin fibers, the skin weakens and becomes susceptible to permanent scars, as an old rubber band tends to lose its elasticity.

The devices commonly used to treat stretch marks problems have recently gained popularity in the consumer market. But they are complex to use and have a high cost. Some are invasive and, in turn, are accompanied by a long-term treatment that the patient can't pay.

In this article, a device that uses electro stimulation as the primary technique to reduce stretch marks will be shown. In Section II, one will focus on explaining the design of the electro stimulator. The electroestimulator is composed of electronic devices at low cost, the frequency ranges that are of reference to know what parameters to use and get the desired result will be defined. Different types explained waves and their effects on muscle and nerve. Section III will debate on the use of electro stimulator and treatment time. In Section IV, the results obtained with the electro stimulator efficiency and to reduce streaks in the different parts of the body where they are located will be explained. The conclusions derived through research to finalize detailed in Section V, and also mentioned recommendations to consider and take into account.

II. Methodology

The flow chart used for the development of the project is shown in Fig. 1. The flow chart begins with a processing and control stage that concludes in a test stage.

Concerning the processing and control of the signals, the placement of electrodes in the area of the body that has stretch marks is performed. Then the electrical signals produce the vibration of the electrodes, and finally, the signals are regulated by using the frequency table of impulses (Table I). The tests were performed on patients affected with stretch marks, obtaining favorable results.

A. Circuit Design

Muscular electrostimulation is a technique that involves the application of electrical pulses by using controlled electrical current to the induction of muscle contraction [2]. The pulses are generated on a device that applies electrodes on the skin near to muscles that are intended to stimulate. They mimic the action impulses from the central nervous system, causing muscle contraction.

The electrodes generally adhere to the skin. EEM is a form of electrotherapy or muscle training [3]. Several authors cite it as a complementary technique for the treatment of muscle; there are numerous studies published on the matter. It was found that this technique could be used as therapy for people with particular rehabilitation-related problems or muscle tissue or muscle problems such as stretch marks disorders. Currently, muscle electro stimulation used for recovery purposes centers or as a form of muscle training.

This circuit design seeks to generate high voltage pulse voltage and low current that is not detrimental to the sensitivity of the human body so that the muscles can contract. The circuit works with 220v in the primary of the transformer, and this so that the output of the transformer we reduced 6.9 or 12-volt alternating voltage, the diodes are responsible for rectifying and smoothing the electrolyte capacitors. The primary function of the circuit of "Fig. 2" is to generate pulses of short duration that are perceived through the electrodes, which, when connected to the body, can contract the muscle.

We design a circuit with the following components:

- 1 LM556
- 1 audio transformer 110v / 12v to 100 milliamps (or less better between 60 to 100miliamperios)
- Resistors 1k ohm, 1k ohm, 390 ohms, 330 ohms, 33k, 220hm (3), 50k
- Capacitors 100 Nano farad, 10 microfarads
- 3 led
- 5k potentiometers, 1M, 10k
- Transistors 2N2222, 2N3053
- Two 1N4004 diodes

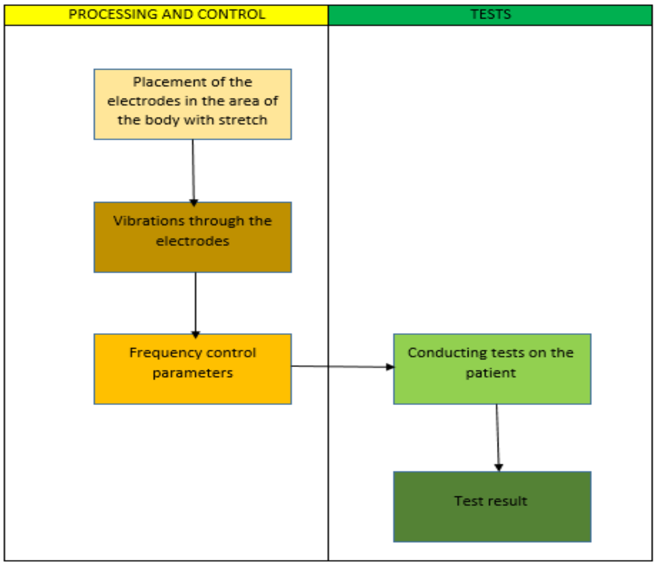

PROCESSING AND CONTROL
TESTS
Placement of the electrodes in the area of the body with stretch
Vibrations through the electrodes
Frequency control parameters
Conducting tests on the patient
Test result

Fig. 1. Diagram Flow.

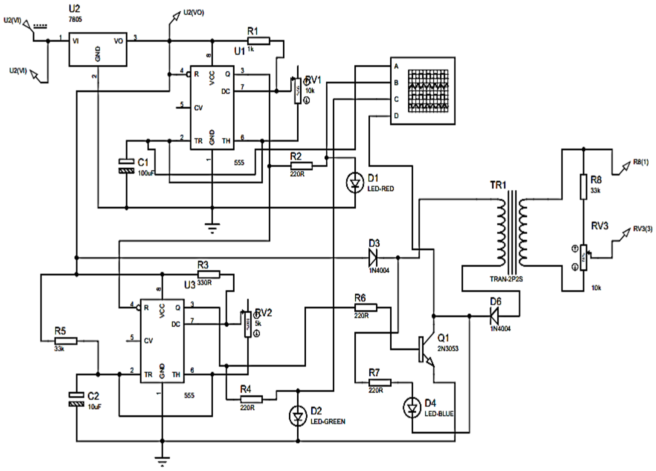


Fig. 2. Electro Stimulator Circuit.

In "Fig. 3", we design the circuit using the Proteus Design Suite, software of electronic design automation, and this to create the tracks on the circuit.

In "Fig. 4", we can see a final prototype 3D, how will be the circuit once it is implemented with electronic components.

B. *Operating Principle*

The critical aspect of electro stimulation lies in knowing in detail the program that you want to work because its effects can be varied [4]. In this sense, if you're looking to lose weight, you should know that the program to follow will be the type of stretch marks, as is the one who can help in this task. This specific program improves the ability to metabolize fats in your body and pass through the anterior muscle to the patient's skin. Also, other sections that one should not overlook, when resorting to the sessions with devices electro stimulation, is that usually, this method is only able to stimulate a muscle group at a time. This situation implies a significant dedication and time consuming to exercise full musculature.

In "Fig. 5", we have an operation diagram which indicates the work to conduct the electro once the electrodes are connected to the body area affected by stretch marks.

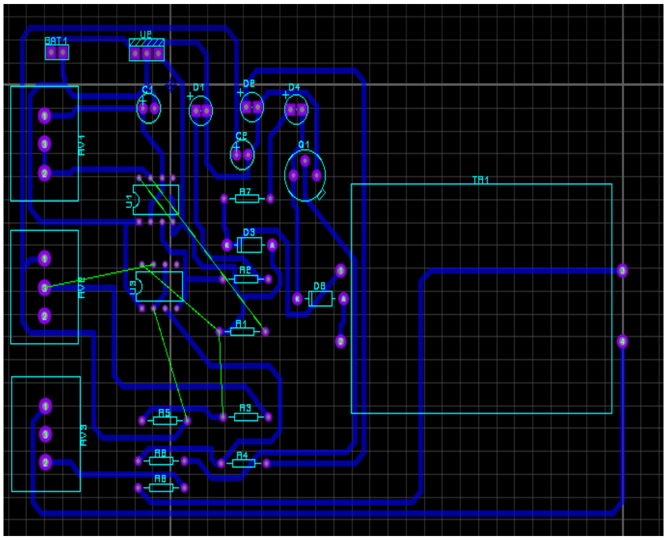


Fig. 3. Track Circuit Design.

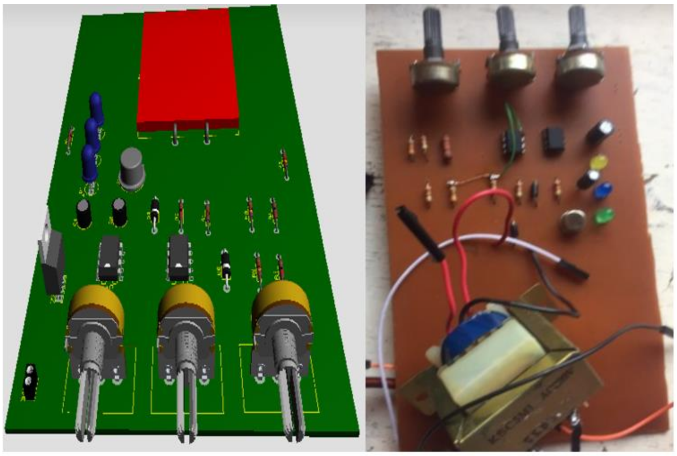


Fig. 4. 3D Simulation Circuit and Implementation of Circuit.

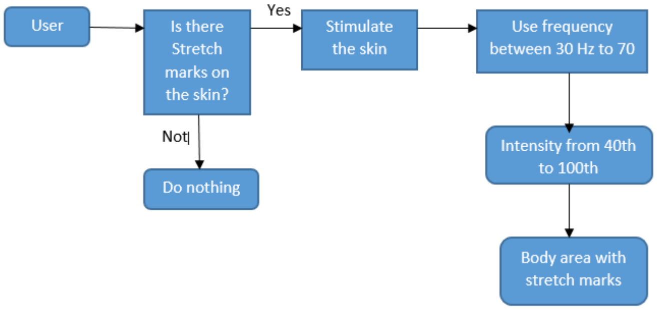

User
Is there Stretch marks on the skin?
Yes
Stimulate the skin
Use frequency between 30 Hz to 70
Not|
Do nothing
Intensity from 40th to 100th
Body area with stretch marks

Fig. 5. Operating Diagram.

C. Electro Stimulation Signal

The signal to produce muscle stimulation would be achieved with electric waves, and thus a reaction in the affected area would be performed. One must apply the electrical parameters that will be used. It must be minimum energy, intensity, and duration, and one must locate the characteristics of the optimal flows and comply with its fundamental law. The most advanced electro stimulator has already programmed the Hz, depending on the frequency. It has the appropriate terminology to improve sports performance (explosive strength, strength, strength resistance, hypertrophy). The search for an aesthetic improvement (muscle firmness, toning, and bodybuilding) on functional recovery and improvement of the quality of life (back pain, neck pain, active recovery, relaxation, drainage) [5]. The electro stimulator of the device recreates the signals that are sent through electrodes in pulses and in response. The muscle reacts with a contraction. The tissue cannot distinguish if the command comes from outside the brain.

D. Pulse Frequency

Is the number of pulses which are repeated at a given time is measured in Hertz? In the table, we set the frequency range that the stimulator should work. While the frequency is higher, the greater the strength and power. Typically, one speaks of three types of fiber types working in the following frequencies:

a) Slow fibers: His frequency begins from the 10 Hz, and peaks at 33 Hz should be clear that always stimulates both slow fibers.

b) Mixed fibers: His tetanization starts at 20 Hz and ending at 50 Hz, frequencies used to improve strength.

c) Fast fibers: Their frequency starts at 33 Hz and ends at 66 Hz However, these are given for a sedentary person values; Sportsmen higher frequencies are used.

E. Electrodes EMS

The electrodes are generally metal plates used as a driver responsible for making contact with a circuit sector non-metallic type, but electrolyte, as the human body (Table I). EMS means electrical muscle stimulation is a technique that promotes muscle contractions by applying electrical impulses. EMS is commonly used for medical and physiotherapy and sport complement treatments; therefore, therapeutic purposes, sports or aesthetic, anti-cellulite, and drainage [6].

TABLE. I. IMPULSE FREQUENCY RANGE

Frequency (Hz)	**Effects of frequencies according to the range used**
1 to 10	Relaxation increased blood flow and segregation of endorphins.
10 to 20	Improving aerobic endurance muscle (muscle oxidative capacity).
20 to 50	Improved muscle tone, muscle definition and muscle firmness (aesthetic effects and early stages of rehabilitation).
40 to 70	Lactic improved capabilities muscle and increase muscle volume.
70 to 120	Improving maximum strength.
90 to 150	Improve explosive, reactive elastic force.

In the body, currents are produced by the movement of ions. The flow of electrons creates a wire for operating a computer.

In "Fig. 6", we show the kind of electrodes we use to the electro stimulator, which are suction electrodes that are useful to lower resistance and increase the electrical conductivity of the skin.

F. Feeding Step

A source that allows the correct operation between blocks mentioned above will be developed. Voltages to be used will be of 9seg and 5volt DC. I know it will require a linear regulator that will allow me to get the 5v. DC. The transistor is using the 2N3053 voltage converter.

G. Waveforms

Currently, the stimulator can work with different types of waves are grouped as follows:

a) Galvanic wave: It is the first form of electric current. It uses low frequency with a particular characteristic that is to maintain a constant voltage at a particular time, as it is observed in “Fig. 7”.

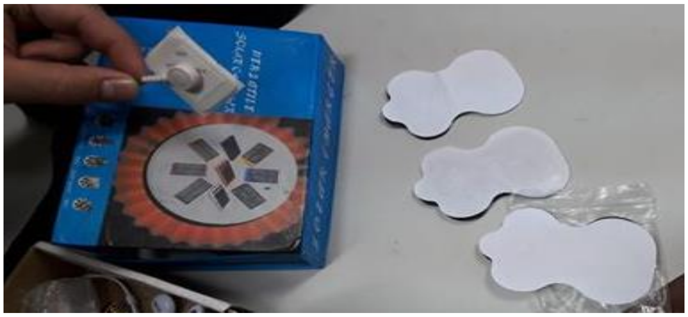


Fig. 6. Suction Cup Electrodes.

b) Faradic wave: The wave is asymmetric, low voltage and works in frequency approximating 50 Hz, which is shown in “Fig 8”. It has a direct application to the muscle to contract it so well innervated.

c) Interferential waves: Two alternating signals having medium frequency shown in “Fig. 9”, varies with slight differences. When the two signals cross a modulated therapeutic current is produced at low frequency, providing a fixed frequency signal with a variable.

d) Monopolar square interrupted galvanic wave: The monophasic waveform is a square pulse shown in “Fig. 10”.

e) Monopolar triangular interrupted galvanic wave: It is a monophasic continuous wave which is shown in Fig. 11. It has a succession of unmodulated pulses and which has a low frequency.

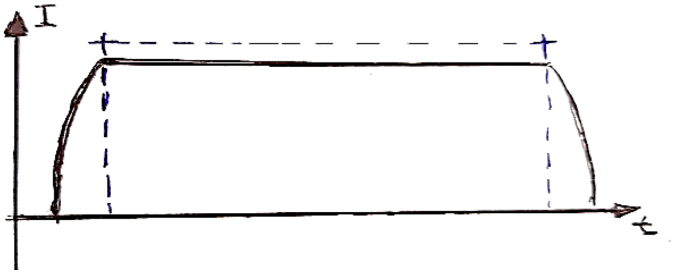


Fig. 7. Galvanic Current.

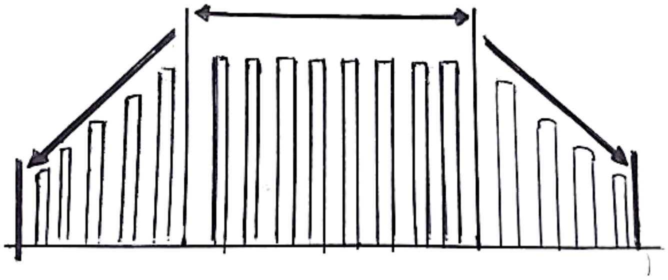


Fig. 8. Faradic Wave Train.

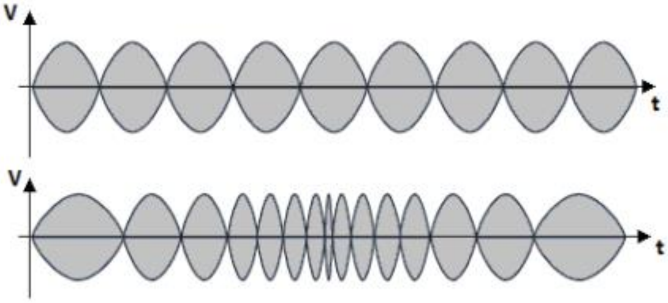


Fig. 9. Interferential Wave Modulation [7].

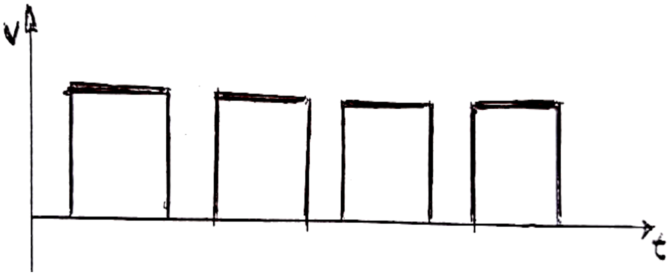


Fig. 10. Monophasic Square Wave.

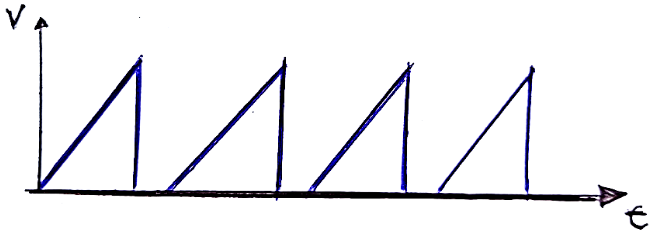


Fig. 11. Triangular Wave Monophasic.

III. Results

Implantation of electrodes in the subject test or patients with this disease of the skin as testing time was that the use was enacted. The idea of reducing or appease the deformation of the skin (stretch marks or whitish lines). That takes place in the process of the same system to the muscles that you get to visualize the skin. One sees that as the results look will be established just said and studied since the beginning of this paper.

The classic question of whether the circuit worked or not, with the time developed a well-implemented circuit that was tested in physics. From there, the research was started to make the paper established, with the time we saw that if This use could reduce the level of stretch marks.

When contracting the muscle makes these same looks oblige to move forcefully. So that they can pass blood, it needs the skin and can see that is bloodying. One needs to see those lines with blood become clogged, and thus the muscle meets the function no longer form whitish lines on the skin, and no more streaks are seen in a long time.

It was tested as the time use also affects the nervous state and can produce Parkinson's too involuntary movement. To avoid this kind of damage long-term type of procedure, which are sessions established by professionals of this branch are available and see how the patient's body behaves.

It will be seen as the time that this will help not only on the part of the biomedical but on the physical side. The blow to the muscle also helps the exercises that can be used in the gym, as this could give movements muscles could see it improve over time.

When the stimulation session is done, it should stimulate isometrically to thereby the muscle is fixed and can prevent the movement that causes contraction.

The electro stimulator that was designed has high levels of muscle contraction, which is suitable for therapies seen.

Table II shows a picture, which will indicate how often you work, time, and the number of sessions is estimated to reduce the use of electro stretch marks shown.

In "Fig. 12", the part of the Axillary nerve shown with problems of stretch marks. To reduce the stretch marks was treated with the electro stimulator, and thus definite reduction of stretch marks was in the "Fig. 13".

In "Fig. 14", connected to the outputs of the electro stimulator, with two oscilloscope channels were shown on channel 1 (yellow signal). Square interrupted galvanic waves of "Fig. 10", and in channel 2 (blue signal) galvanic waves interrupted triangular of "Fig. 11".

TABLE. II. RAME OF REFERENCE FOR ELECTRO STIMULATOR USE IN AREAS OF THE BODY AFFECTED BY STRETCH MARKS

Part of the body affected	**Frequency (Hz)**	**Electro stimulator usage time (minutes)**	**Number of sessions**
Axillary nerve	40-50	10	15
Muscle epitrochlear	50	10	10
Rectus	50	15	15
Oblique muscle	50	10	15
Buttocks	50-60	15	15
Gracilis muscle	50-80	15	15
Vastus muscle	20-50	15	20
Tensor fascia	50	10	15

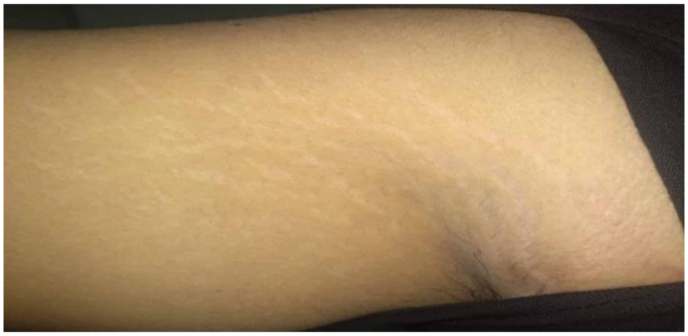


Fig. 12. Stretch Mark in Axillary Nerve.

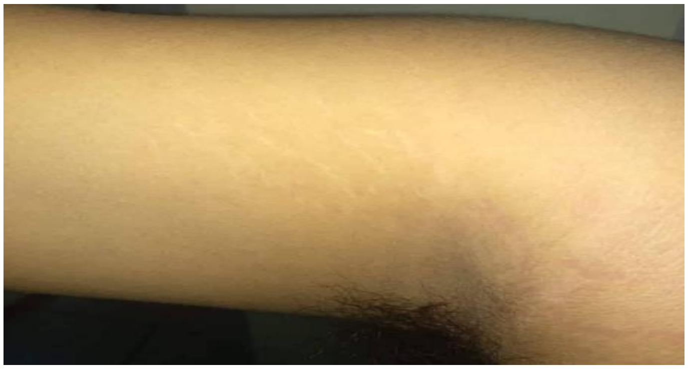


Fig. 13. Reduction Stretch marks in Axillary Nerve with the use of an Electro Stimulator.

IV. DISCUSSION

The classic question of whether or not work the circuit, with the time developing a circuit be implemented that was tested physique. Hence the investigation split to make the paper established with the time one saw that if one could give this use that would reduce the level of stretch marks.

Other of the main drawbacks was the consumer of this product would be too, well thought as time. That using this would be excessive, which had to propose sessions a long time and use series as you see the improvement.

One thought about the disturbance of the aesthetic by which is also used as machine sport in theory if it is true. But the dimension of this paper is to demonstrate by testing that if one can improve and release of this congenital disease that occurs. Dare have too much time for rest or morbidly obese and thus could also see muscles dead, as dead skin by forming these whitish lines.

When the stimulation session is done, it should stimulate isometrically to thereby muscle is fixed and can block the movement that occurs muscle contraction spontaneously.

In the decade of the nineties it is discovered that with the percutaneous application of microloans current with certain pulses mounted to it, it is possible to generate a controlled local inflammatory response that undermines tissue repair, without systematic effect [8]. Currently, our electro stimulator allows a better result in less time through electromagnetic pulses.

Pixelated radiofrequency is based on the fractional unipolar radiofrequency emission to produce a double effect on both, the skin surface and depth [9]. We propose the use of the electro stimulator to produce muscle contractions in the skin.

To help prevent the appearance of stretch marks and varicose veins, and to be able to burn the unnecessary calories for her and the baby, and avoid the overweight of the pregnant woman [10]. Physical activity in pregnant women does not have to demand much physical wear to keep the baby's health in good condition.

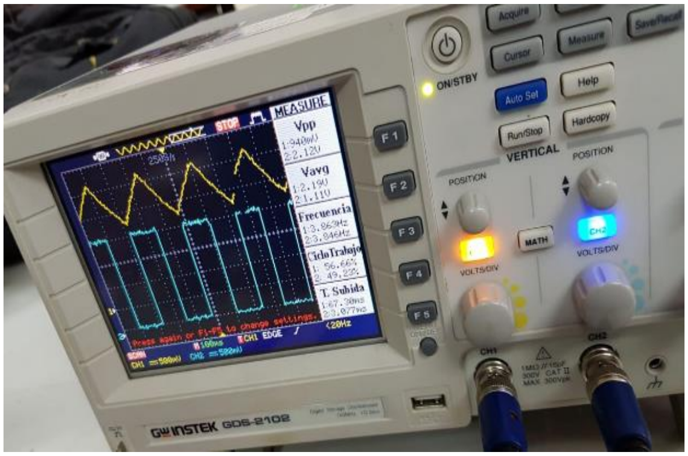


Fig. 14. Signal Square Wave and Triangular through Oscilloscope.

V. CONCLUSIONS

Is to explain everything about the definition, causes, and treatments of different types of stretch marks.

It was designed and implemented to a device of easy use to reduce stretch marks does not discriminate sex and age. The primary function of the electro stimulator is based on improving and toning the resilience of the skin.

It is essential to identify the degree of stretch marks level to give proper treatment with electrostimulation.

To have favorable outcomes with the use of electrostimulation treatment should be supplemented with functional exercises that can stimulate and accelerate the recovery of damaged skin. The idea is to have better muscle contraction as quickly as possible.

The frequency range for the correct use of electrostimulation was investigated to provide special treatment to the patient according to frequencies and stipulated.

The advantage of the electro stimulator is the reduction of stretch marks in the short term. It is also low cost and can be used by the population with small economic resources.

REFERENCES

[1] Teresa O. Rueda, "Securing non-ablative fractional laser 1540nm" in Spain, May 2014 pp 25-26.

[2] Cruz A, "electrotherapy muscle strengthening" in Peru, pp 50-53, June 2018.

[3] Adalbert F. Armijos, "Design and implementation of a prototype of psychomotor early electrostimulation in children with the Down syndrome of the Despertasr Los Angeles center to improve motility in upper extremities" in Ecuador, pp. 20-21, October 2018.

[4] García, R. Marquez, S. Gomez, RI, & Beltran, C, "Efficacy, safety, and cost-effectiveness of transcutaneous electrical nerve stimulation (TENS) for the treatment of skeletal muscle pain chronic" in Seville, pp. 5-7, April 2013.

[5] Pérez Fariñas L. "Treatment of Patient With abdominal obesity as prevention or progression of the metabolic syndrome." Cuban Journal of Physical Medicine and Rehabilitation. 2014; 6 (2-15).

[6] Leandro A. Cajina, Yanira Hernandez and Consuelo C. A. Rivera, "Application methods of electrotherapy in national hospitals and physical rehabilitation centers in the east, the Savior" in El Salvador, pp 41-42, October 2016.

[7] Martin R, "Form of the waves in low and medium frequency currents used in electrotherapy", pp. 35-63, Spain 2014.

[8] Mosquera T. Tatiana, " In vivo evaluation of the cosmetic efficacy of two biostimulation procedures with the application of platelet-rich plasma on stretch marks, to improve the elasticity and firmness of the treated skin" in Ecuador, pp 33-34, November 2017.

[9] Aguero Z. Fatima, Gonzalez B. Lourdes, “Treatment of stretch marks " in Paraguay, pp 86-88, August 2017.

[10] Anchundia M. Jean, “Benefit of physical activity to the pregnant woman from the perspective of nursing." in Ecuador, pp 8-11, August 2018.

Robust Video Content Authentication using Video Binary Pattern and Extreme Learning Machine

Mubbashar Sadddique[1], Khurshid Asghar[2], Tariq Mehmood[3], Muhammad Hussain[4], Zulfiqar Habib[5]
Department of Computer Science, COMSATS, University Islamabad (Lahore Campus), Lahore, 54000, Pakistan[1, 5]
Department of Computer Science, University of Okara, Okara, 56300, Pakistan[2]
Department of Computer Science & Information Technology[3]
Superior University, Lahore, 54000, Pakistan
Department of Computer Science[4]
King Saud University, Riyadh
Saudi Arabia

***Abstract*—Recently, due to easy accessibility of smartphones, digital cameras and other video recording devices, a radical enhancement has been experienced in the field of digital video technology. Digital videos have become very vital in court of law and media (print, electronic and social). On the other hand, a widely-spread availability of Video Editing Tools (VETs) have made video tampering very easy. Detection of this tampering is very important, because it may affect the understanding and interpretation of video contents. Existing techniques used for detection of forgery in video contents can be broadly categorized into active and passive. In this research a passive technique for video tampering detection in spatial domain is proposed. The technique comprises of two phases: 1) Extraction of features with proposed Video Binary Pattern (VBP) descriptor, and 2) Extreme Learning Machine (ELM) based classification. Experimental results on different datasets reveal that the proposed technique achieved accuracy 98.47%.**

Keywords—Video forgery; spatial video forgery; passive forgery detection; Video Binary Pattern (VBP); feature extraction

I. Introduction

In these days, digital video making has become very handy with the accessibility of video recording gadgets such as smartphones and digital cameras [2, 1]. These videos are an important part of our daily routine and also an important source of information. Digital videos present some of the most convincing documentary evidence to establish the truthfulness or falsehood of an issue under consideration, which is acceptable both inside and outside the court of law.

A few years back, digital videos were considered reliable proof, but a widespread availability as well as accessibility of easy-to-use video editing tools (VETs) such as (Pinnacle Studio 20 Ultimate, Adobe Premier Pro, Lightworks and Cinelerra, etc.) [21, 19], have negated this fact. Even a novice user can alter the contents of digital videos in such a manner that it is not possible to distinguish between the original and forged contents of a video with the naked eye. On one hand, video editing is a very useful and important tool for manipulating video scenes in film industry. On the other hand, it enables to forge video contents to distort the evidences for a court and propaganda on social, print and electronic media Therefore, authenticity of a video is a key issue when it is presented in a court as a prof of a crime [12].

Digital video forgery techniques are categorized into temporal, spatial, and spatio-temporal. In spatial category, digital videos are forged by changing contents within the frame(s) which modifies visual information. The object is taken from one location of a video frame and inserted on another place in the same frame or in other frame after some alteration [17]. This category consists of upscale-crop [15], copy-move [18] and splicing video forgery [5].

Temporal tampering (forgery) is done by removing, duplicating or inserting the number of frames from / in a digital video. Both object and frame level forgery is done in spatio-temporal category. Existing tampering detection techniques in digital videos are divided into active and passive. Active techniques need pre-embedded data such as watermarking, digital signatures, etc., whereas passive techniques do not depend on any pre-embedded information. Passive techniques are also called blind techniques. Fig. 1 shows categorization of video forgery techniques.

Various passive techniques are proposed to detect spatial video forgery which are not equally efficient for different datasets, static and moving objects. In this research, a robust video content authentication technique is presented. This paper is structured as follows: Section II explains related work; Section III describes a step-by-step research methodology used for development of proposed technique; datasets and performance evaluation parameters are described in Section IV; and experimental work is presented in Section V. Conclusion and future directions are presented in Section VI in the end of this paper.

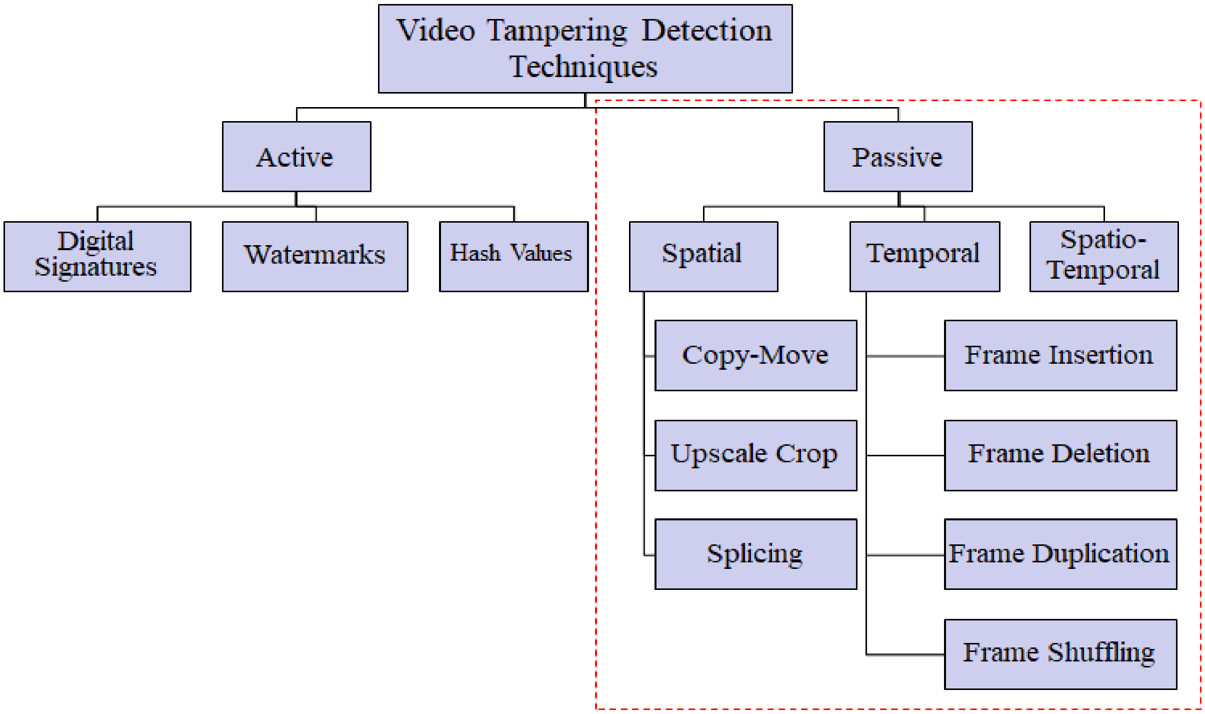


Fig. 1. Categorization of Video Tampering Techniques.

II. Related Work

Existing techniques for tampering detection in spatial domain can generally be classified as: 1) Statistical Based, 2) Compression Based, 3) Texture Based, and 4) Noise Based. Fig. 2 describes types of video forgery detection techniques. Each technique is briefly described in the following discussion.

A. Methods based on Statistical Features

Texture, tone, and context are always present in any frame (image). A texture is an important property, which is disturbed during the process of video forgery. The statistical features are used for representation of the texture [29]. Many researchers used statistical features for detection of object-based video forgery [10, 3, 2]. Singh et al. [2] exploited DCT and correlation coefficients to detect the duplicated regions. The method achieved accuracy 99.5% and 96.6% for detection of duplicated frames and regions respectively. This method cannot detect less number of duplicated frames and is not able to detect small duplicated regions. Richao et al. [10] used statistical features to detect forgery. Authors calculated the first four moments of the wavelet and average gradient of each color component. SVM classifier is applied for classification between original and forged video. Twenty videos with resolution 320 x 240 were used for an experiment. The accuracy and area under the curve (AUC) are 95% and 0.948 respectively. The receiver operating characteristic (ROC) curve showed a classification result of 85.45%. The results are tested on a limited dataset and not experimented on different compression ratios. Su et al. [3] proposed an algorithm based on exponential Fourier moments (EFMs) to detect the duplicate region and adaptive parameter based compression tracking algorithm is used to localize the tampered region. The detection accuracy 93.1% is achieved.

B. Methods based on Compression

Su et al. [9] proposed a new algorithm based on compressive sensing for video forgery detection. When the moving-objects deleted from a video frames, the traces (lines, edges, and corners) around the object are also altered. The compressive sensing is employed to perceive this tampering. The K- Singular Value Decomposition (K-SVD) was used to obtain and analyze difference between frames. The detection results of each frame were combined to obtain result. This technique has more compliance in problem solving and is more easy to use as compared to another similar technique proposed in [23]. However, the proposed system does not perform well for a very small deleted foreground.

A technique of forgery localizing in MPEG-2 videos was presented by Labartino et al. [14]. The proposed method first discovers twice intra-coded frames and then applies a double quantization analysis based on MPEG-2. The technique exploited the properties of MPEG-2 coding. This novel technique encoded by utilizing P-frames for analysis of video to apply double quantization. A well-known video dataset was used to carry out experiments having varied scenes having 720x576 resolution.

C. Methods based on Texture

Tamura texture features are exploited for detection of copy-move video forgery by Liao et al. [13], these features are utilized to localize the copy-move tampering. The method was verified on a dataset having 10 videos, which are captured with fixed and moving camera. The resolution of videos was 640 × 480 and a frame rate of 25 to 30 frames per second (fps). The results reveal 99.96% precision, which is comparatively higher than previous appropriate research. However, computation time of the method is much higher.

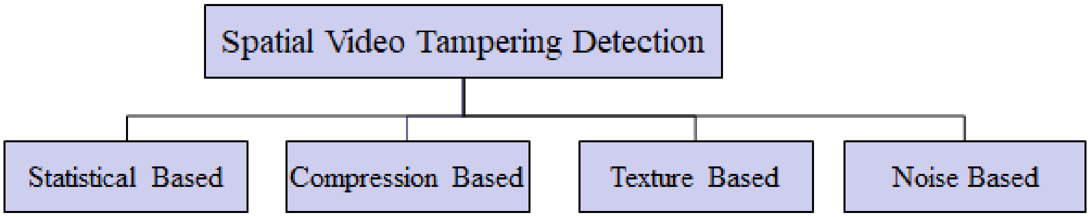


Fig. 2. Types of Techniques for Video Tampering Detection in Spatial Domain.

Subramanyam et al. [17] detected the spatial forgery by using compression and Histogram of Oriented Gradients (HOG) features. In this study, the authors used 6000 frames from 15 different videos with tampered regions of size 40x40, 60x60 and 80x80 in the same and different frames. Detection accuracy (DA) is 80%, 94% and 89% for 40x40, 60x60 and 80x80 blocks, respectively. This algorithm detected spatial forgery very well, but the training and testing are performed on a very limited dataset. The algorithm fails to detect tampering when different post-processing operations (rotation and scaling) are exercised to forge the regions and cannot localize the forged regions.

D. Methods based on Noise Characterisics

Noise variations between an authentic and tampered video sequence are exploited to detect forgery. The photon shot noise in digital camera was exploited by Kobayashi et al. [24]. The test was performed on gray scale videos recorded on 30 fps and 640×480 resolution, Huff-yuv, a lossless compression codec was used for compression. The experiments were performed on videos recorded from static scene only.

A bottom-up approach based on noise correlation was used by Hsu et. al. [25] to locate the forged / in-painted regions of a video. The noise residual is calculated by subtracting the noise free video frames and original frames. Wavelet de-noising filter is used to obtain noise free video frames. Then, every video frame was divided into non-overlapping $N \times N$ blocks. Then correlation of the noise residual was calculated between successive frames. Lastly, forged blocks are located by analysis of statistical features. Content dependency of the noise residual made correlation feature unstable for applications for moving cameras.

III. Proposed Methodology

In this section, a robust video content authentication technique in spatial domain is presented. The proposed methodology consists of two stages (see Fig. 3). (1) Feature extraction, and (2) classification based on ELM**, e**ach stage is described in the following discussion.

A. Feature Extraction

Feature extraction, a first step of proposed methodology, plays a very crucial role to distinguish whether a video is authentic or forged. These isolate important characteristics in the video. These unique characteristics of an authentic video do not exist in the forged video. Existing techniques for feature extraction have a high computational cost. Moreover, normally, features are extracted frame-wise from a video by taking these frames as static images. Therefore, no temporal correlation exists between these features, which are very important in video forgery detection. Importance of features demands the need of a descriptor that extracts features, which are, not only discriminate, but also has temporal correlation. These features are helpful in classification of original and forged videos. A descriptor is proposed to calculate the features named as Video Binary Pattern (VBP) and processes of a proposed descriptor are described below.

- As a first step, digital video is converted into '*N*' number of slices. The number '*N*' depends on '*Height (H)*' of video frames. For example, a digital video consisting of frames with image height $H = 240$ is converted into 240 slices. By converting video into slices, temporal correlation between the frames is accessed. Fig. 4 describes the process for slicing of digital video.
- Feature extraction from such a large number of slices is time consuming. To minimize the computation time, the number of slices are reduced in a manner that they do not affect classification, whatever the resolution of the video. We extract average of these slices using following formula:

$$X = \frac{N\ number\ of\ Slices}{Y}, \tag{1}$$

where *Y* is 10 in our experiments.

$$\text{Average Slice, } W = \frac{Sum\ of\ X\ Number\ of\ Slices}{X}, \tag{2}$$

- The desired features are then extracted from these average slices using Local Binary Pattern (LBP) and are represented as:

  F_1 = LBP (S_1), F_2= LBP (S_2),….. F_n = LBP (S_N), where '*N*' and 'n' are indices of slice numbers and feature vectors, respectively.

- *n* Features of all slices are concatenated to make a final vector (*V*) which is represented as:

$$V = \text{Concatenate}\ (F_1, F_2, F_3, \ldots, F_n) \tag{3}$$

B. Classification based on ELM

Classification is a process of categorizing groups of data based on similarities. The classification models attempt to extract some decisions from observed data. When an input is given to classification model (trained model), it will predict and categorize the given value into one or more groups. For example, when a feature vector is given as input to the trained model, it will label as authentic or forged. Different models are available for classification, for example, support vector machine (SVM) [28], ELM [20], Decision Trees and Naïve Bayes [11], etc. The proposed descriptor was trained and tested using ELM. It has superiority in computational speed as compared with other machine learning algorithms such as Decision Trees and Naïve Bayes etc. The proposed descriptor was trained and tested using ELM. It has a single hidden layer feedforward neural network and has superiority in computational speed as compared with other machine learning algorithms and, being a mathematical model, its implementation is easy. We have experimented different classifiers such as J48 which is used to generate decision trees, Naïve Bayes based on Bayes' theorem with strong independence assumptions between features, Multiclass classifier [27] which classifies instances into one of three or more classes, SVM a supervised learning model, simple ELM [26] feedforward neural networks based classifier and Kernel ELM [22] where the number of neurons is decided by the classifiers itself. The proposed technique finds best result with kernel ELM using RBF kernel.

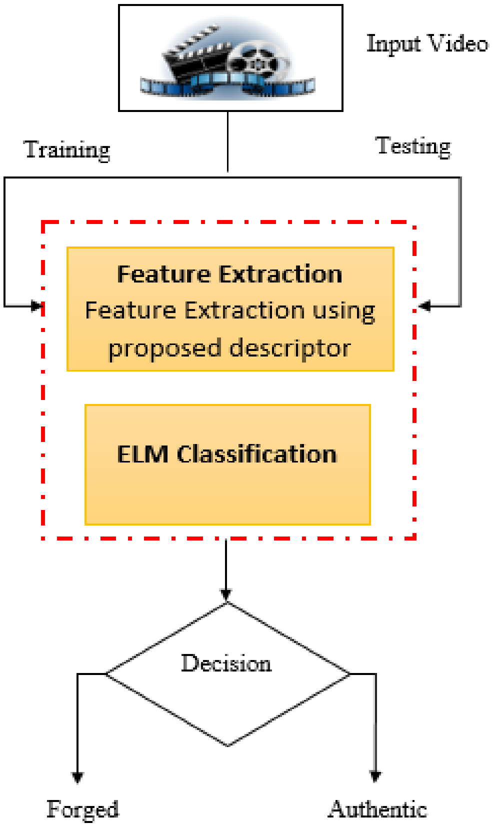


Fig. 3. Proposed Methodology.

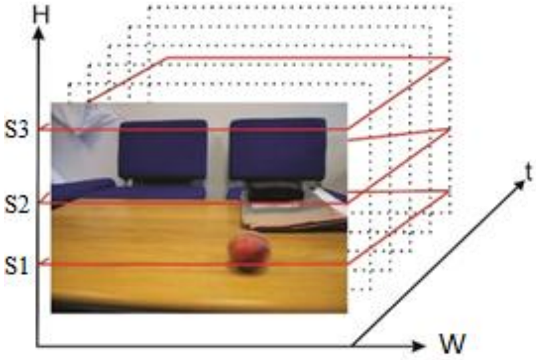


Fig. 4. Model for Slicing of Digital Video.

IV. Experimental Results

The performance of proposed technique was measured on different datasets taken from Hsu et al. [25], Bestaigini et al. [16], Ariddizone and Mazzola [8], and Sanjary et al. [6] which are summarized in Table I. Sample frames taken from authentic and forged video sequences of these datasets are shown in Fig. 5 and 6 which describe different objects removed, inserted or duplicated to forge digital videos. The elephant object is inserted at frame No. 250 in forged frame (see Fig. 5(f)) to forge the video content. Similarly the car object is copied and inserted in the same video. The forged frames (50 and 375) of forged video are shown in Fig. 6(d) and 6(f).

The performance of proposed technique is evaluated by different evaluation parameters same as discussed by Richao et al. [10]. A frame is labelled as true positive (TP), if a forged frame is also classified as forged. Otherwise, if the frame is classified as an original, then it is labelled as false positive (FP). A frame is labelled as true negative (TN), if an original frame is also classified as original. Otherwise, the frame is labelled as false negative (FN). Accuracy (AR) is the ratio of sum of TPs and TNs to the sum of TPs, TNs, FNs and FPs. True positive rate (TPR) performance parameters is a ratio of total TPs to the sum of total TPs and FPs.

False positive rate (FPR) is a ratio of total FPs to the sum of total TNs and FNs. AR, TPR and FPR are represented by the following equations:

$$AR = \frac{TP+TN}{TP+TN+FN+FP} \quad (4)$$

$$TPR = \frac{TP}{TP+FP} \quad (5)$$

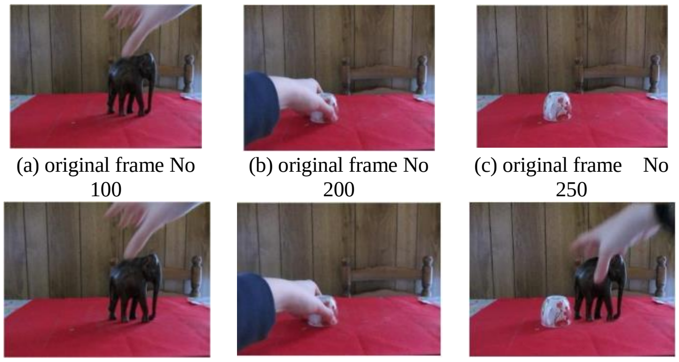
$$FPR = \frac{FP}{TN+FN} \quad (6)$$

Experiments were performed on Intel ® Core ™ i5-2400 CPU @ 3.10GHz, 64-bit Windows operating system with 4GB RAM using MATLAB version R2018a. Experimental results of proposed technique tested on DS1 to DS5 are shown in Table II. The results reveal that the proposed technique demonstrates higher accuracy rate 98.45% using ELM Kernel Classifier on data set DS5 whereas lowest on data set DS2. The accuracy rate is lowest on DS2 because this dataset has less video sequences (10 videos). The classifier, on less data learned specific pattern but generalization of this pattern was not possible. Table III demonstrates effects of different classifiers on accuracy rate using proposed techniques. The best accuracy is achieved using Kernel ELM classifier. Kernel ELM classifier decides by itself the number of neurons to be set.

The performance of our proposed technique is compared with two other existing state-of-the-art techniques proposed by Chen et al. [4] and Richao et al. [10] in spatial domain. Chen et al. claimed accuracy of 99.9% whereas Richao et al. claimed 97.36% accuracy on limited video dataset. We implemented these techniques and obtained results on dataset DS5, because their datasets are not publicly available. The comparison is presented in Table IV. It can be seen from Table IV that our proposed technique outperforms the other techniques with AR = 98.47%, TPR = 98.50% and FPR = 98.37%. Moreover, performance of both the techniques reduced significantly on the proposed datasets.

TABLE. I. DETAIL OF DATA SETS USED FOR EXPERIMENTS

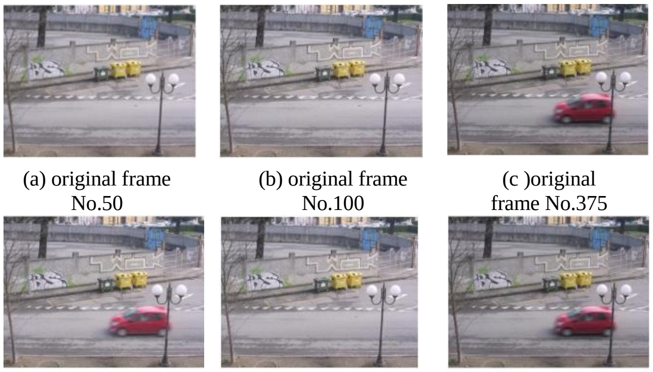
Datasets	**Types of forgery**	**Authentic**	**Forged**	**Frame Rate**	**Static/Moving Camera**	**Resolution**	**Format**			**Length (Sec,)**
							AVI	**MP4**	**WMV**	
DS1 [16]	Copy-move	10	10	30	Static	320x240	20	-	-	7-19
DS2 [25]	Splicing	14	6	30	Moving	720X480	15	-	5	3-17
DS3 [8]	Copy-move & Splicing	6	121	25-30	Static & Moving	768X576	101	26	-	2-16
DS4 [6]	Copy-move & Splicing	20	20	29	Static & Moving	720x1280	-	40	-	14-15
DS5 (DS1+DS2+DS3+DS4)	Variable	50	157	25-30		Variable	136	66	5	2-19



(a) original frame No 100 (b) original frame No 200 (c) original frame No 250

(d) forged frame No 100 (e) forged frame No 200 (f) forged frame No 250

Fig. 5. Sample of Frames (a)-(c) Taken from Original Videos of DS1 and (d)-(f) from Forged Videos of DS1 [16].



(a) original frame No.50 (b) original frame No.100 (c)original frame No.375

(d) forged frame No. 50 (e) forged frame No. 100 (f) forged frame No. 375

Fig. 6. Sample of Frames (a)-(c) Taken from Original Video and (d)-(f) Taken from Forged Video [16].

TABLE. II. PERFORMANCE OF PROPOSED TECHNIQUE USING ELM

Dataset	**AR (%)**	**TPR (%)**	**FPR (%)**
DS1	96.23	97.51	12.23
DS2	92.56	96.20	30.12
DS3	93.83	92.35	25.67
DS4	97.47	96.50	21.32
DS5	98.47	97.10	14.23

TABLE. III. EFFECTS OF DIFFERENT CLASSIFIERS ON ACCURACY

Classifier	**AR (%)**	**TPR (%)**	**FPR %)**
J48	82.35	82.4	43.34
Naïve Bayes	66.45	66.74	36.12
SVM	60.89	60.30	32.67
Multiclass classifier	83.08	83.34	40.56
Simple ELM	97.45	96.10	19.45
Kernel ELM	98.47	97.10	14.23

TABLE. IV. COMPARISON OF PROPOSED TECHNIQUE WITH STATE OF ART

Methods	**AR (%)**	**TPR (%)**	**FPR (%)**
Proposed Technique	98.47	97.10	14.23
Method proposed by Chen et al. [7]	72.67	68.31	29.66
Method proposed by Richao et al. [10]	63.6	59.42	38.77

V. CONCLUSION

Detection of forgery in a video is a challenging task, because it substantially affects content of the video. In this paper, we proposed a two-stage technique (feature extraction and classification) for forgery detection in spatial domain. For features extraction a descriptor Video Binary Pattern (VBP) is proposed to extract features from average slices of videos and ELM classifier is used for detection and classification of video forgery. Experimental results on different datasets reveal that the proposed technique achieved accuracy rate 98.47% using ELM classifier. The technique is also robust to different formats and variety of datasets.

Further research will also be required to enhance the accuracy through cross dataset validation, which is important for reliable and real time applications.

REFERENCES

[1] M. Zampoglou, F. Markatopoulou, G. Mercier, D. Touska, E. Apostolidis, S. Papadopoulos, R. Cozien, I. Patras, V. Mezaris, and I. Kompatsiaris, "Detecting Tampered Videos with Multimedia Forensics and Deep Learning", in International Conference on Multimedia Modeling, 2019, pp. 374-386, 10.1007/978-3-030-05710-7_31.

[2] G. Singh and K. Singh, "Video frame and region duplication forgery detection based on correlation coefficient and coefficient of variation", Multimedia Tools and Applications, vol. 78, pp. 11527-11562, 2019, 10.1007/s11042-018-6585-1.

[3] L. Su, C. Li, Y. Lai, and J. Yang, "A Fast Forgery Detection Algorithm Based on Exponential-Fourier Moments for Video Region Duplication", IEEE Transactions on Multimedia, vol. 20, pp. 825-840, 2018

[4] S. Chen, S. Tan, B. Li, and J. Huang, "Automatic detection of object-based forgery in advanced video", IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, pp. 2138-2151, 2016, doi.org/10.1109/tcsvt.2015.2473436

[5] K. Asghar, Z. Habib, and M. Hussain, "Copy-move and splicing image forgery detection and localization techniques: a review", Australian Journal of Forensic Sciences, pp. 1-27, 2016, doi.org/10.1080/00450618.2016.1153711

[6] O. I. Al-Sanjary, A. A. Ahmed, and G. Sulong, "Development of a video tampering dataset for forensic investigation", Forensic science international, vol. 266, pp. 565-572, 2016

[7] S. Chen, S. Tan, B. Li, and J. Huang, "Automatic Detection of Object-based Forgery in Advanced Video", 2015

[8] E. Ardizzone and G. Mazzola, "A tool to support the creation of datasets of tampered videos", in International Conference on Image Analysis and Processing, 2015, pp. 665-675, doi.org/10.1007/978-3-319-23234-8_61

[9] L. Su, T. Huang, and J. Yang, "A video forgery detection algorithm based on compressive sensing", Multimedia Tools and Applications, vol. 74, pp. 1-16, 2014, 10.1007/s11042-014-1915-4.

[10] C. Richao, Y. Gaobo, and Z. Ningbo, "Detection of object-based manipulation by the statistical features of object contour", Forensic science international, vol. 236, pp. 164-169, 2014, doi.org/10.1016/j.forsciint.2013.12.022

[11] T. R. Patil and S. Sherekar, "Performance analysis of Naive Bayes and J48 classification algorithm for data classification", International journal of computer science and applications, vol. 6, pp. 256-261, 2013

[12] Z. Parmar and S. Upadhyay, "A Review on Video/Image Authentication and Temper Detection Techniques", International Journal of Computer Applications, vol. 63, 2013

[13] S.-Y. Liao and T.-Q. Huang, "Video copy-move forgery detection and localization based on Tamura texture features", in Image and Signal Processing (CISP), 2013 6th International Congress on, Hangzhou, China, 2013, pp. 864-868

[14] D. Labartino, T. Bianchi, A. De Rosa, M. Fontani, D. Vazquez-Padin, A. Piva, and M. Barni, "Localization of forgeries in MPEG-2 video through GOP size and DQ analysis", in Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on, 95 Pula (CA), Italy, 2013, pp. 494-499

[15] D.-K. Hyun, S.-J. Ryu, H.-Y. Lee, and H.-K. Lee, "Detection of upscale-crop and partial manipulation in surveillance video based on sensor pattern noise", Sensors, vol. 13, pp. 12605-12631, 2013

[16] P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Local tampering detection in video sequences", in Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on, Pula (CA), Italy, 2013, pp. 488-493, 10.1109/MMSP.2013.6659337.

[17] A. Subramanyam and S. Emmanuel, "Video forgery detection using HOG features and compression properties", in Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on, Banff Center Banff, AB, Canada, 2012, pp. 89-94, doi.org/10.1109/mmsp.2013.6659337

[18] V. S. Pujari and M. Sohani, "A Comparative Analysis On Copy Move Forgery Detection Using Frequency Domain Techniques", International Journal of Global Technology Initiatives, vol. 1, pp. E104-E111, 2012

[19] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensics", APSIPA Transactions on Signal and Information Processing, vol. 1, p. e2, 2012, doi.org/10.1017/atsip.2012.2

[20] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification", IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, pp. 513-529, 2012

[21] A. Rocha, W. Scheirer, T. Boult, and S. Goldenstein, "Vision of the unseen: Current trends and challenges in digital image and video forensics", ACM Comput. Surv., vol. 43, pp. 1-42, 2011, 10.1145/1978802.1978805.

[22] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification", IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, pp. 513-529, 2011

[23] J. Zhang, Y. Su, and M. Zhang, "Exposing digital video forgery by ghost shadow artifact", in Proceedings of the First ACM workshop on Multimedia in forensics, 2009, pp. 49-54

[24] M. Kobayashi, T. Okabe, and Y. Sato, "Detecting video forgeries based on noise characteristics," in Advances in Image and Video Technology. vol. 5414, ed: Springer, 2009, pp. 306-317.

[25] C.-C. Hsu, T.-Y. Hung, C.-W. Lin, and C.-T. Hsu, "Video forgery detection using correlation of noise residue", in Multimedia Signal Processing, 2008 IEEE 10th Workshop on, Queensland, Australia, 2008, pp. 170-174, doi.org/10.1109/mmsp.2008.4665069

[26] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications", Neurocomputing, vol. 70, pp. 489-501, 2006

[27] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression", Bioinformatics, vol. 20, pp. 2429-2437, 2004

[28] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers", Neural processing letters, vol. 9, pp. 293-300, 1999

[29] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification", Systems, Man and Cybernetics, IEEE Transactions on, pp. 610-621, 1973.

Automated Greenhouses for the Reduction of the Cost of the Family Basket in the District of Villa El Salvador-Perú

Pedro Romero Huaroto[1], Abraham Casanova Robles[2], Nicolh Antony Ciriaco Susanibar[3], Avid Roman-Gonzalez[4]
Electronics and Telecommunication Department[1, 2, 3]
Aerospace Sciences and Health Research Laboratory (INCAS- Lab)[4]
Universidad Nacional Tecnológica de Lima Sur, Lima, Peru[1, 2, 3, 4]

***Abstract*—Today, the cost of the family basket is gradually increasing, not only globally but also in our country. This increase includes the demand for vegetables and fresh vegetables that allow people to improve their quality of life. Also, this consists of the search for a healthier and more natural diet with the help of existing technologies. Against this, we propose the implementation of an automated greenhouse, with sensors and actuators that allow controlling a microclimate for correct and efficient development of vegetables. With this proposal, we obtain a saving of 50% equivalent to 8.00 dollars, concerning the planting of lettuce against market prices, thus achieving a reduction in the family basket.**

***Keywords*—Germination; protocol i2c; hummus; atmega**

I. Introduction

The automated greenhouses are gaining rapid reception, thanks to this; it is possible to plant non-native plants within these environments. Nowadays, the tremendous technological advance makes possible the recreation or artificial implementation of natural habitats and processes. This situation has, as a consequence, a benefit and a higher yield of the products that are elaborated in the said process [1].

At present, the creation of greenhouses for the development and growth of plants or vegetables favors different ecosystems and reuse of it [2]. Several studies indicate that the production of plants in these protected and isolated structures have higher growth and creation that is reflected in the energy, water, and economic savings where one aims [3]. As they adapt to the necessary conditions of climate change, which leads to optimizing the conditions of crop growth. Management is improved at the time of providing efficient irrigation and increases efficiency in production [4].

According to the National Institute of Statistics and Informatics of Peru (INEI), one knows that the cost of the family basket is gradually increasing. In 2015 to 2016, the increase was 0.041%, and from 2017 to 2018, the increase was drastic at 0.93% affects families in our country [5].

Faced with this problematic situation, the objective of this work is to implement an automated greenhouse. This greenhouse works automatically and generates decisions by executing corrective actions, to adjust the climatic parameters of the variables such as temperature, humidity, and humidity of the land.

The importance of our idea is to reduce the cost of the family basket and improve people life quality. Also, the use of recycling in the implementation process is to take care of the environment and minimize the costs of assembling the automated greenhouse. Likewise, one focuses on the goal of self-sustainable development for the benefit of families in Villa El Salvador.

The idea is to give a second option to people who are in a situation of poverty. The idea is that people can have the capacity to generate their essential consumption foods and in a sustained manner over time, with a minimum of investment. We will take advantage in a specific part of the eco-friendly means of the District Villa el Salvador, like the humidity that oscillates between 80% to 90%, that is very favorable for the sowing of certain products [6].

There are several studies regarding the reduction of the family basket; one shows the following antecedents to have a current vision of this problem.

In [7] the authors corroborate the importance of automated greenhouses in the study of pathogens or diseases, which could affect the growth of the tomato crop in the process of its development. Likewise, they are stimulated with different climatic changes to see the diversity of diseases that affect tomato.

The author in [8] proposes a technological alternative that helps to mitigate the problem of the cost of the family basket through a start-up aimed at municipalities. Likewise, the creation of the CanastAPP is a web platform where the products of first need at a lower market cost.

In [9], the author makes a situational diagnosis, with the purpose of particular the poorest neighborhoods of its community. This diagnosis is focused on organic urban agriculture, to obtain a well to be with natural products and to reduce the expenses in the purchases of the familiar basket. Likewise, it conducts market studies to determine the demand for organic products, and quantify said demand to show these products in urban agriculture.

In [10], the Peruvian state carries out social programs to support the economy and the family basket. These are the milk glass program, the complementary feeding program-PCA, national program, direct support to the poorest and national plan of school feeding Qali Warma. All these programs are state initiatives, support people who have fewer resources and improve their quality of life.

In [11], the authors are looking to improve the quality and productivity of lettuce consumption in the national market, through a large-scale hydroponic greenhouse. Due to the current deficiency in vegetable growers, this research is motivated to offer a quality product and think about the final consumer.

In [12], it is proposed to improve the production of vegetables in high Andean areas in Cusco - Peru to 3330 masl. Through an automated greenhouse prototype taking advantage of the materials of the city for the construction of the same, obtaining a significant saving of 10 dollars per square meters.

This work continues as follows: Section II shows the methodology that is developed to reduce the cost of the family basket in the Villa El Salvador District. In Section III, the authors present the results obtained. Finally, in Section IV, one can see the discussion and conclusion.

II. Methodology

A. Implementation of the Automated Greenhouse

For the construction and implementation of the computerized greenhouse, one searched for a relatively small area to perform the tests with dimensions of 2 meters long by 1.2 meters wide. The implementation is followed by plastic supports to make the structure for the aspects of the greenhouse. Then, proceeded to make the electrical wiring for the power supply of the sensors and actuators, so that they can be connected to the analog reading pins using cable UTP category 5 of 8 threads. The DTH 11, MQ2 sensors and humidity sensor were placed strategically in the greenhouse, to obtain better readings. Likewise, the LCD screen with the I2C adapter was placed in the front, so that anyone can see the real-time data of the variables counted. The analog reading terminals are redirected to a point in common so that it reaches the central controller, which is the Arduino mega 2560. One uses the relay module for the actuators to separate the different suppliers to create a stable system. In Fig. 1, one can appreciate the flow of implementation and results.

Finally, one closes the greenhouse hermetically with transparent plastic. The idea is to let's pass the solar brightness and this way to realize the proper growth. One can see in Fig. 2 the scheme of implementation.

It also has to get adequate land for planting that meets high standards of nutrients. To obtain a product of better quality and at a low cost, this land must be fertilized with hummus and fertilizers, which help the growth of the ideally. It also seeks proper oxygenation to meet the essential standards in the planting of lettuce.

Two tests will be carried out to see the development of the vegetable, and parameters such as the quality and time of its development will be seen. The first test will be done in the greenhouse where one will plant seeds of silk lettuce. This variety of lettuce takes between 8 to 12 days to give the first shoots, with favorable climatic conditions. One wants to reduce germination to fewer days to be able to have it before possible the product. Since it is isolated and controlled by the greenhouse, it will climatize the environment with the purpose of the correct development of the silk lettuce. It will give adequate watering to the land that must include moisture values between 60% and 80% that are desirable for the development of the plant. Likewise, it will provide an adequate temperature with ranges between 18 and 20 degrees Celsius and an ambient humidity necessary for the rapid growth of the plant. The idea is to control any incidence of gases that may affect the development of the vegetable and be able to have it in normal conditions.

The second test, one will take an area of 1.2 meters long by 1.2 meters wide. In this space, one will make planting of lettuce, but it will be exposed to environmental conditions, we will put the necessary measures to isolate and enclose it. It will be in an environment without any care in terms of temperature, soil moisture, humidity, and gases. The idea is to be able to visualize the growth in environmental conditions, and that takes so long to make the germination of the first shoots of lettuce to be able to make a comparison. If indeed, the automated greenhouse generates a better product, in a shorter time and a higher quality. The greenhouse will be careful under favorable conditions that would give us a saving in the use of electric power, greater efficiency in water use, and the economic part. This situation reflects basically the cost of the family basket. It would be an excellent start to recreate this system in each house of our district to generate a better quality of life.

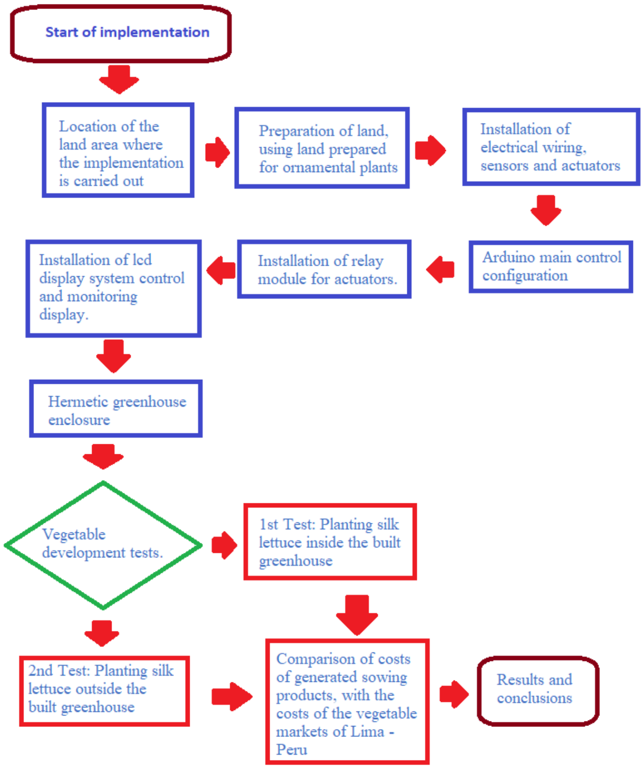

Start of implementation
Location of the land area where the implementation is carried out
Preparation of land, using land prepared for ornamental plants
Installation of electrical wiring, sensors and actuators
Arduino main control configuration
Installation of relay module for actuators.
Installation of lcd display system control and monitoring display.
Hermetic greenhouse enclosure
Vegetable development tests.
1st Test: Planting silk lettuce inside the built greenhouse
2nd Test: Planting silk lettuce outside the built greenhouse
Comparison of costs of generated sowing products, with the costs of the vegetable markets of Lima - Peru
Results and conclusions

Fig. 1. Flowchart of the Implementation Process and Results.

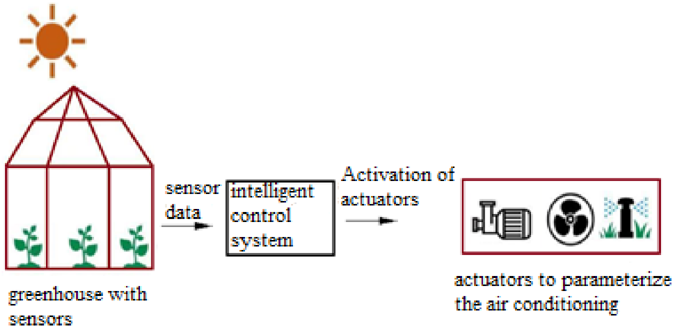

sensor data
intelligent control system
Activation of actuators
greenhouse with sensors
actuators to parameterize the air conditioning

Fig. 2. Automated Implementation Schemes.

B. Climate Variables

1) Temperature: The temperature is an essential variable in the development of the germination and the growth of the silk lettuce. It must be controlled efficiently to have a healthy crop and above all of better quality. Then, one will manage specific ranges of temperature for the development of our lettuce, and this can be seen in Table I.

2) Humidity: The humidity is an important variable, and the necessary care must be given. Since an excess of humidity in the environment can damage the growth of the plant with different factors, as well as fungi or rot on the leaves. The environment should be humidities between 60% and 80% for adequate development of lettuce, as seen in Table II.

3) Soil moisture: The soil moisture similar to the humidity should be between 60% and 80%. If in case it is less than this amount will delay the growth of the lettuce. Likewise, the irrigation that must be done is a small amount of watering of water to avoid rottenness in the neck of the stem, and in Table III, one can see these values.

4) Carbon dioxide: To control the variable carbon dioxide in our automated greenhouse, one uses an HS-1 that is a portable CO2 meter with a range of 0 to 2500 ppm. The lettuce needs, for its cultivation, the range of 900 ppm to 1600 ppm for the correct photosynthesis. Otherwise, the 1600 ppm barrier will pass the air extraction system that is activated to level the CO2 levels has been basal.

5) Mineral content: For the development of the research, one uses the horiba laqua meters. They can measure levels of ph, nitrate, potassium, calcium, and more minerals that help the development of the plant. One monitor in real-time to correct any deficiency. In Table IV one can see those values.

6) Light: The variable light in the research of lettuce planting is significant for the development of the crop and its photosynthesis. In the same way, the greenhouse is designed to be able to receive natural light through the transparent cover that is on the roof.

C. Pests and Diseases

Every vegetable has pests that affect the development of the crop. Likewise, one is in the district of Villa El Salvador that presents humidity relative to 80% and conducting soil studies. One finds a pest that directly affects the planting of lettuce and is the warm gray that develops in this type of weather. These worms attacks feeding on the roots or on the shoots that are closest to the ground. To eradicate them one use a mixture of pyrethroid with water and through the water pump and drip irrigation. The insecticide is dispersed reducing the pest and controlling it.

1) Alternaria dauci: It is a disease that occurs in lettuce when it is exposed to high levels of humidity. Fungi that are difficult to detect with the naked eye appear are like dark spots on the leaves. These fungi control them efficiently and reduce high levels of humidity, activates the air extraction system. This situation is in order to pass through a salt filter to internally reduce the humidity of the greenhouse and control the disease.

D. Programming and Operation

When talking about the programming and its operation, one addresses to the platform of assembly of the program. In this case, Arduino will be used. First, one needs the necessary libraries; in this case, we will use two essential libraries that are: include "DHT.h" and include <LiquidCrystal_I2C.h>. These libraries allow better control of the temperature and humidity sensor DTH 11 and the liquid crystal i2c libraries. It is to reduce the use of pins and use the i2c communication protocol, for improved transmission of data in real-time to a screen LCD.

TABLE. I. TEMPERATURE °C OF THE LETTUCE

PHASES OF CULTIVATION			
	OPTIMAL	MINIMUM	MAXIMUM
GERMINATION	18°C-20°C	14°C	28°C
INCREASE	16°C-19°C(day) 12°C-16°C(night)	11°C	28°C
FRUIT	16°C-19°C(day) 13°C-16°C(night)	14°C	28°C

TABLE. II. ENVIRONMENTAL HUMIDITY OF THE LETTUCE

PHASES OF CULTIVATION			
	OPTIMAL	MINIMUM	MAXIMUM
GERMINATION	60%-80%	50%	95%
INCREASE	60%-80%	45%	85%
FRUIT	60%-80%	45%	85%

TABLE. III. HUMIDITY OF EARTH H%

PHASES OF CULTIVATION			
	OPTIMAL	MINIMUM	MAXIMUM
GERMINATION	60%-80%	40%	98%
INCREASE	60%-80%	40%	85%

TABLE. IV. GROUND MINERALS

MINERALS FOR CROP		
	MINIMUM	MAXIMUM
PH	5.8 ph	7.2 ph
CALCIUM	15 gr	17gr
POTASSIUM NITRATE	2 kg	3 kg

The DTH 11 sensor has the functionality to read the temperature and humidity variables in real-time, and the control flow for temperature can be seen in Fig. 3.

The MQ2 sensor has the functionality of being able to read the variable of propane gas in a basal state that is around 50G to 110G. If it exceeds the basic parameters that one has programmed in the 180G command lines, which may influence or affect the greenhouse ecosystem, the air extraction system will be activated in order to stabilize the automated greenhouse as seen in Fig. 4.

Earth moisture sensor, like the other sensors, will record the soil moisture of our greenhouse, to see if the proportion of humidity that should comprise values higher than 70%. Otherwise will activate the water pump to generate drip irrigation and the control flow can be seen in Fig. 5. In Fig. 6, one can see the general diagram block.

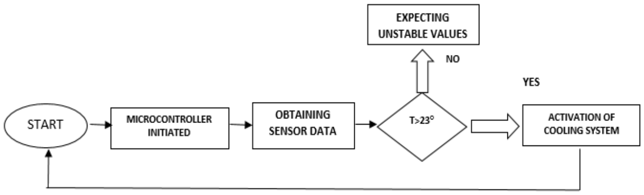


Fig. 3. Temperature Control Flow.

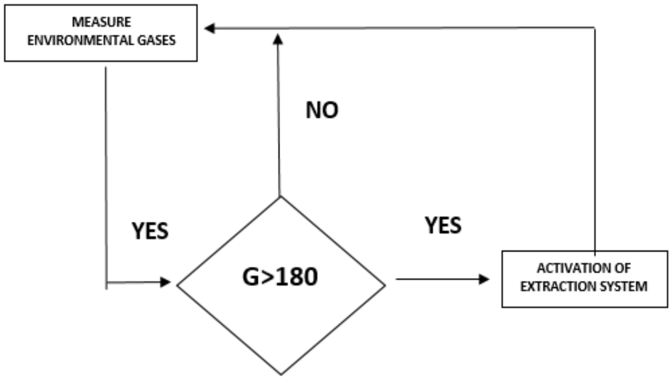


Fig. 4. Ambient Gas Control Flow.

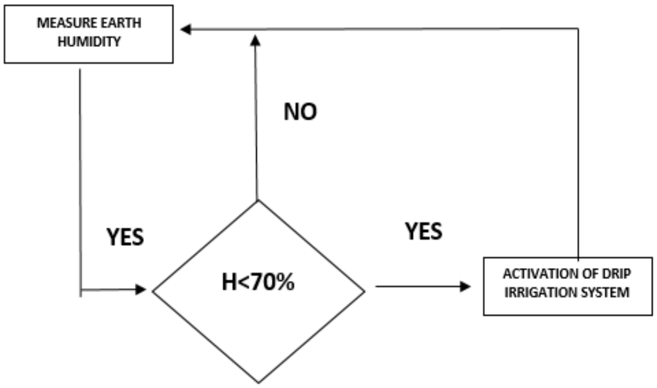


Fig. 5. Ground Moisture Control Flow.

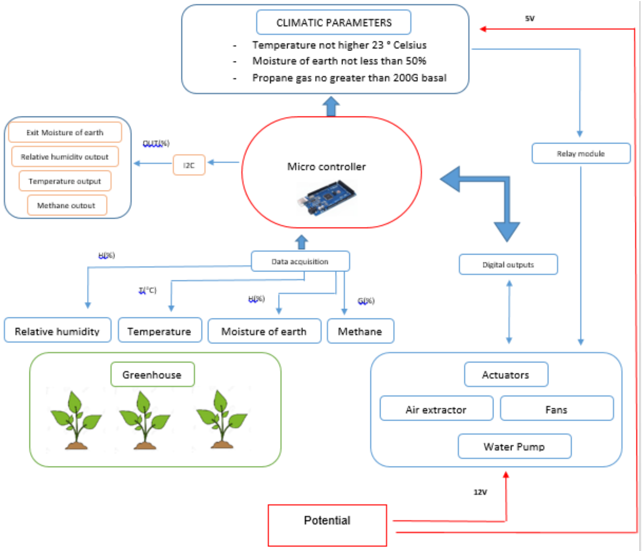


Fig. 6. Automated Greenhouse Control Schemes.

III. Discussion

Based on the findings, one can say that the cost of the family basket can be reduced by implementing low-cost automated greenhouses in the Villa El Salvador District. Likewise, the objectives of sustainable development are met. Poverty, improve health and well-being in terms of the quality of life of the population.

As for the results obtained, one finds itself in total disagreement with the research "Organic Orchard Gardens in the San José neighborhood" [8] since, in the period of one year, 744 dollars was spent to obtain 372 kilos of lettuce for 87 families. This situation would be 4.27 kilos for a cost of 8.55 dollars per family, which unlike was spent 5 dollars on the planting of the vegetable and one has 45 lettuces with a projection to increase. Likewise, the process used to farming is mechanical, both irrigation and ventilation, which brings a slowdown in the growth of the plants and not an automated process like ours.

Our result correlates with the construction of a Start-up focused on saving money in the purchase of products from the family basket [7]. Since one is going in only one direction, which is to reduce the cost of the family basket, it would be integrated in an appropriate way to plant the fast-growing vegetables. The start-up would be used to acquire the missing products to have an adequate nutritional supplement. This situation would have significant savings in terms of the economic. The improvement of the family basket would also be improved with the help of Peruvian state policies with social programs, giving this problem a plus [10].

One agrees with the results given the research of greenhouses in high Andean areas. Since our study gives us the economic savings values are very similar in our case 8 dollars and the greenhouses high Andean areas 10 dollars. All these significant savings is to improve the family basket and take advantage of productivity [12].

IV. Results

With a period of 20 days, one obtained 40 shoots of lettuce with a projection to an increase in the passing of days. Surveys were conducted in the Santa Rosa market and the Santa Anita Producers market to know the real cost of the lettuce. The idea is to make the respective analysis for the economic savings in its research, and one can see the results in Table V.

TABLE. V. Analysis and Comparative Results of Prices of Lettuce in the Markets of the District of Villa el Salvador

Criteria	Automated Greenhouse	Producers Market of Santa Anita	Santa Rosa Market of Villa el Salvador
Surveys to Sellers of vegetables	1	7	6
Price lettuce unit	$/. 0.20	$/. 0.30	$/. 0.40
Savings per unit of lettuce	$/. 0.10 - $/. 0.20	$/. 0.00	$/. 0.00

As can be seen in the results, there is a significant saving when comparing the two most important markets in the districts of Lima and Villa el Salvador, in terms of the cost of lettuce. While in the greenhouse the cost per unit of lettuce is $0.20, in the market the price is higher. If the lettuce were sold at the market price that is $0.40 each, we would obtain a saving of $0.20 per unit multiplied by the 40 lettuces obtained would be $8.00.

One also has significant savings in the implementation of the automated greenhouse, because we use recyclable materials to reduce the costs in the assembly of the structure. One also built and implemented the entire project, which was significant in the total completion greenhouse.

V. Conclusion

The research carried out provides an intelligent automated greenhouse system. The controls different environmental parameters, improving the productivity and increase of the vegetables (as seen in the results). Consequently, a reduction of the cost of the family basket in the District of Villa el Savior.

Another essential objective is the reuse of recycling in the process of implementation, to lower the costs of assembly of the automated greenhouse. Likewise, we focus on the aim of sustainable development, which is responsible for production and consumption to take care of the environment.

TABLE. VI. Greenhouse Implementation Costs

AUTOMATED GREENHOUSE	PRICE
HARDWARE ELECTRONIC (arduino, sensors, actuators, chargers, hs1, horiba laqua)	$/. 65
HARDWARE MATERIALS (talk structure, cover, belts, cables)	$/. 30
SOFTWARE	$/. 15
PEOPLEWARE WORK	$/. 30

Regarding the actions to be taken, the idea is to present this research to the people of the District of Villa El Salvador. With this situation, they get involved with the technology. Parallel to this, take an environmental awareness and see other options to improve the quality of community life, as seen in Table VI.

References

[1] Felipe A. Yevilao Cuevas, “Automated Temperature Control, Humidity, Irrigation by Misting for lettuce scale cultivation”, 2018, pp 28-30.

[2] Hernan Alarcon .L, Geyni Arias .V, Javier Diaz and David Soto .V, “Design of a control and automation system for temperature, ground humidity and relative humidity to optimize crop yield under cover in CORHUILA” 2017, pp 49-51.

[3] Gurban, Eugen Horatiu; Andreescu, Gheorghe-Daniel “Environmental Engineering & Management Journal” (2018): vol 17: pp. 399-416.

[4] Nicolosi, Volpe, y Messineo “An innovative adaptive control system to regulate microclimatic conditions in a greenhouse.” 2017, pp 722.

[5] INEI “Idicators of price of the Economy” 2018, pp 63-71.

[6] INEI “Yearbook of environmental statistics” 2018 pp 63-66.

[7] Vandana Rangrao .H “An automated climate control system for greenhouse use deep learning for tomato crops” 2018, pp 94-96.

[8] Carlos Eduardo Pabón .G “Construction of a Startup focused on the Money Saving in the Acquisition of Products of the Family Basket” 2019, pp 20-39.

[9] Fany Asqui Yuquilema“Feasibility study for the implementation of family organic gardens in the San neighborhood José Del Vínculo, parish of Sangolquí, Rumiñahui canton” 2018, pp 36-43.

[10] Office of the Comptroller General of the Republic of Peru "report of the control services of the social programs in charge of the state" 2015, pp. 33-37.

[11] Chirinos Centes Adolfo, Herrera Lagos Renzo“ Implementación de un invernadero a escala para la creacion de una empresa productora de lechuga hidroponica en Lima Metropolitana” 2016, pp 16-35.

[12] Zambria Pacheco Pedro .G “Invernaderos sostenibles para la producción de hortalizas en zonas alto andinas del Cusco t” 2019, pp 97-104.

Learning Analytics Framework for Adaptive E-learning System to Monitor the Learner’s Activities

Salma EL Janati[1], Abdelilah Maach[2], Driss El Ghanami[3]
LRIE Laboratory, Mohammadia School of Engineers (EMI)
Mohammed V University, Rabat, Morocco

***Abstract*—The adaptive e-learning system (AE-LS) research has long focused on the learner model and learning activities to personalize the learner’s experience. However, there are many unresolved issues that make it difficult for trainee teachers to obtain appropriate information about the learner's behavior. The evolution of the Learning Analytics (LA) offers new possibilities to solve problems of AE-LS. In this paper, we proposed a Business intelligence framework for AE-LS to monitor and manage the performance of the learner more effectively. The suggested architecture of the ALS proposes a data warehouse model that responds to these problems. It defines specifics measures and dimensions, which helps teachers and educational administrators to evaluate and analyze the learner’s activities. By analyzing these interactions, the adaptive e-learning analytic system (AE-LAS) has the potential to provide a predictive view of upcoming challenges. These predictions are used to evaluate the adaptation of the content presentation and improve the performance of the learning process.**

Keywords—e-Learning; adaptive e-learning system; learner model; learning analytics; business intelligence; data warehouse; content presentation

I. Introduction

In the last few years, there has been a growing interest in the adaptive e-learning system (AE-LS). It is a new approach that makes an e-Learning system more effective by adapting the content presentation to the learner in accordance with their preferences, knowledge, and behavior. The aim of AE-LS is to provide the appropriate information to the right learner at the right time. It is based on a learner model, which used to adapt learner’s interactions of the e-Learning system according to their needs [1]. However, AE-LS does not cover all learning aspects since it does not provide teachers and designers with the adequate tools, which allow them to monitor and access to learner's activities [2].

Many teachers, and designers expend enormous amounts of effort to design their learning monitoring system for the e-learning system to maximize the value of learner‘s interactions between tutors, learners, and content. These systems allow them to access and summary information of learning activities. These interactions comprise number of visited pages, the time spent on each page content, and the date of the first and last access to the application. This information is important to have an idea about learners’ activities, but not enough to analyse their behavior and perform their evolution. Therefore, the teacher has a serious problem to extract information and use it in a useful way.

To address this issue, a new business intelligence framework, named adaptive e-learning analytic system (AE-LAS), has been suggested. In this approach, we suggest an analytic framework for AE-LS using business intelligence tools. This approach consists of the extraction of data from the AE-LS and the transformation of these data to store it in data warehouse tools, in order to visualize the learner’s activities. The result of this visualization will be used to predict learners’ preferences and improve the performance of the learning process. This paper is structured as follows: The first section presents a literature review. The second section describes the suggested architecture of the AE-LS and details their feathers. The next section presents the case study and the result of the suggested approach. Finally, we conclude with conclusion and future work.

II. Literature Review

AE-LS and Learning Analytics (LA) are two complementary approaches to monitor learner’s activities. In the review, the decrease in usage of open data sets, providing of AE-LS is likely a reflection of the privacy issues that arise with each LA application. To note also that the number of LA applications that use data from AE-LS has significantly decreased as compared to early academic analytics studies, which focused more on improving learner retention, graduation rates and the adaptation of the course content to the learner.

A. *Adaptive e-Learning System (AE-LS)*

AE-LS is a set of technologies and approaches that combined to give learners an adequate content meeting their needs [3]. In adaptive e-learning approach, each learner has different characteristics, which not suitable for another type of learners [4]. AE-LS covers a variety of systems: E-learning Platform [3], Learning Management Systems (LMS) [5], Intelligent Tutoring Systems (ITS) [6], Adaptive Hypermedia Systems (AHS) [7], and Dynamic Adaptive Hypermedia Systems (DAHS) [8]. The adaptation in AE-LS is divided into four elements: adaptive content aggregation, adaptive presentation, adaptive navigation, and adaptive collaboration support as is presented in Fig. 1. Adaptation content aggregation consists of providing the learner with several content depending on his/her background, preferences, learning style, and his/her level [9]. While, the adaptive presentation attempted to adapt the content presentation to the learner’s need by introducing the adequate visual presentation (text, audio, video, etc.) [10].

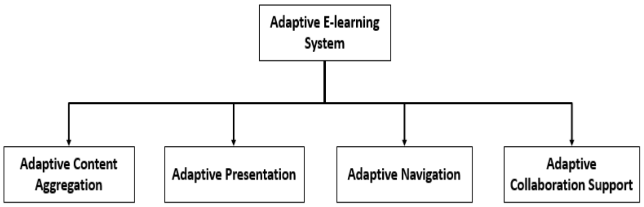

Adaptive E-learning System
Adaptive Content Aggregation
Adaptive Presentation
Adaptive Navigation
Adaptive Collaboration Support

Fig. 1. Adaptive E-learning System Structure.

The adaptation of navigation is based on the links presented in the pages, these links allow the navigation between pages, personalize view in the page, and facilitate the adaptation process [11]. The last adaptation of collaboration support allows the communication between learners and the collaborative application using network-based communication [12].

These systems have suffered from a lack of analytical reporting tools using to track the learner performance, evaluate the visual content presentation, and improve the adaptation of the system. Most adaptive e-learning system applied potential Learning Analytics tools either to tell learners what to do next by automatically matching learning resources to the individual needs of the learner and recommending them different learning entities, based on their preferences of content, presentation, and navigation.

B. Learning Analytics (LA)in AE-LS

Learning Analytics (LA) has strong roots in a variety of fields. It is a multi-disciplinary field, which combines business intelligence, artificial intelligence, machine learning, data mining, and visualization.

LA has attracted a great deal of attention in AE-LS. It uses varied sources of educational data. These data sources of educational data fall into five categories: Open educational Data sets, E-learning Platform, Learning Management Systems (LMS), Intelligent Tutoring Systems (ITS), and Hypermedia Systems (HS). LA is oriented to be used by learners, teachers, intelligent tutors, intelligent mentors, instructors, and administrators. It is dedicated to measure, perform and analyse data based on the context of the learner in order to evaluate their learning process [13]. Learning Analytics study researches to use intelligent methods, procedures and models to improve the learning process for learners, teachers and instructors point of view. These researches are based on information science, including computer science, psychology, and pedagogy [6].

LA consists of four major steps organized as an iterative cycle: Data Collection, Pre-processing, Analytic Action, and Post-processing [14]. In the first step, LA collects data from various e-learning system and heterogeneous data sources to explore it in the next step. The Pre-processing allows the transformation data into a suitable format that can be used for a particular Learning Analytics method. The third step is based on the result of preparation and transformation data applied in the pre-processing. It includes analysis, adaptation, and visualization data. The last steep of the LA process allows compiling a new data, determining new indicators and metrics and modifying the variable in order to continue the improvement of the analytics system.

The main purpose of LA is to develop a model that attempts to predict learner's behaviors and future performance, based on their current activities and accomplishments. This model tracks learner activities and generate reports in order to support decision-making by the institutors and teachers. This predictive model can then be used to provide an active intervention for learner who may need additional assistance or suffer from impairments.

To achieve these purposes LA uses different methods and techniques such as: Business Intelligence (BI), Statistics, Data Mining (DM), Information Visualization (IV), and Social Network analysis (SNA).

1) Business Intelligence (BI): In the literature, Business Intelligence is defined as a set of methods and concepts used to improve business decision making by a support system. It includes architectures, databases, methodologies, and applications [15].

BI is the set of tools for analyzing data for decision making using data extraction, transformation and loading (ETL) tools to consolidate and integrate data into Data Warehouse (DWH) or OLAP (On-Line Analytical Processing) [16]. DWH is considered the foundation of BI. Some research considers that DWH is the database of the system where data is stored and consolidate [17]. Others researches consider that the DWH is a business intelligence platform [18]. In the education field, BI is used to explore the learner's activities under different perspectives, dashboard, and the report of the visualization. It is used to improve teaching and monitor learner's activities using data log files of e-learning system [19].

2) Data Mining (DM): Data Mining is the process of analyzing and performing data from database, Web, text, image in several perspectives and summarizing it into useful information. This process allows to find correlations or patterns among dozens of fields in large databases and Data warehouses [20].

Several studies have demonstrated that Data Mining techniques could successfully be incorporated into e-learning. It can be used to resolve classification and prediction problems in e-learning. The most famous techniques used to solve these problems are: fuzzy logic networks, artificial neural networks, evolutionary computation, and association rules graphs and trees [21].

3) Statistics: Most existing e-learning systems implement reporting tools that provide basic statistics of learners' interactions and e-learning system. Statistics in e-learning generate a simple statistical operation such as: standard deviation, evolution, average, and sum. For examples the number of visits course, the number of connections by day, time spent on the page, and frequency of learner's replies [22].

4) Information Visualization (IV): Information visualization techniques provide a visualization dashboard for teachers and learners, so that they no longer need to drive blind. It represents the results of LA methods in a user-friendly visual form might in order to facilitate the interpretation and the analysis of the e-learning data [23]. There are different Information visual techniques such as: Bar chart, Box plot, Distribution plot, Combo chart, Gauge, Histogram, KPI, and Treemap, used to represent the information an understandable format.

5) Social Network Analysis (SNA): Social network analysis techniques have been applied in different Learning Analytics tasks. It is the quantitative study of the relationships between learner and e-learning system. Social network analysis is modeled by a graph G = (V, E), where V is the set of nodes representing actors, and E is a set of edges, representing a certain type of linkage between actors [24]. Fig. 2 presents the learning analytics process.

Moreover, LA is still in the early stages of research. However, LA tools suffered from several limitations using with e-learning system. In one hand, learners learn through the Internet, so their learning process cannot be collected by e-learning systems and teachers cannot know the progress of the course, the manner which learner learns, and learner preferences. All of these will lead to the gap between students and teachers. On the other hand, the pedagogical content is static and the teacher's assignment and content given to every learner is the same content. While, learner shaves different background and the knowledge structure is dynamic.

For this reason, we should analyse learning habits, characteristics, knowledge structure, and improve the visualization of the analysis results to teachers to be more intelligible. Instead, more contextual information is needed to be captured in heterogeneous media and graphics. In order to solve these problems, we inspired by approaches previously presented, especially the Business Intelligence approach to build a new dimensional modeling of data warehousing to visualize the learner‘s interaction and reconstruct the adequate learning process of learners. The most famous learning analytics monitoring tools used in e-learning are presented in Table I.

In this section, we present the architecture of the suggested learning analytics system used for AE-LS. The proposed architecture consists of three layers: ETL Process Layer, Data Warehousing Layer, and Restitution Layer. Fig. 3 presents the proposed learning analytics system for AE-LS.

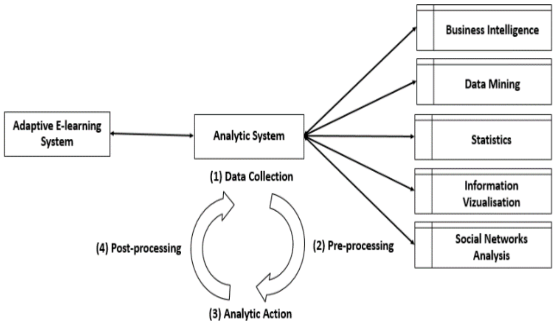


Fig. 2. Learning Analytics Process.

TABLE. I. FAMOUS LEARNING ANALYTICS MONITORING TOOLS USED IN E-LEARNING

AE-LS	LA Tools	Description
E-learning Platform	MATEP [25]	It is a monitoring web interface tool used to help teachers to analyze and visualize the learner data. This tool uses information from log files and add the learner‘s contextual information.
	Monitoring Virtual Classroom [26]	It is a virtual monitoring tool used to follow learner in e-learning system with similar visions of classical classroom.
	Sinergo/ColAT [27]	It is a monitoring tool that offers interpretative views of the activity developed by learners in a group learning collaborative environment. It integrates the information from log files with contextual information.
LMS	CourseVis [19]	It is a learning analytics tool for tracking the learner data and extract it through the web log files of the LMS server to help teachers to follow learners and identify their needs.
	GISMO [28]	It is a learner monitoring tool, designed to extract data from Moodle log files and represented it in graphics. These representations allowed teachers to perform learner's behaviors.
ITS	Educational Data Mining Tool [29]	It is a monitoring tutorial tool, which use tutor logs directly into database to extract data, in order to summarize the learner's interactions.
	TADA-ED [30]	It is a tool, which integrates various visualization and data mining facilities to help teachers and instructors in the learning process.

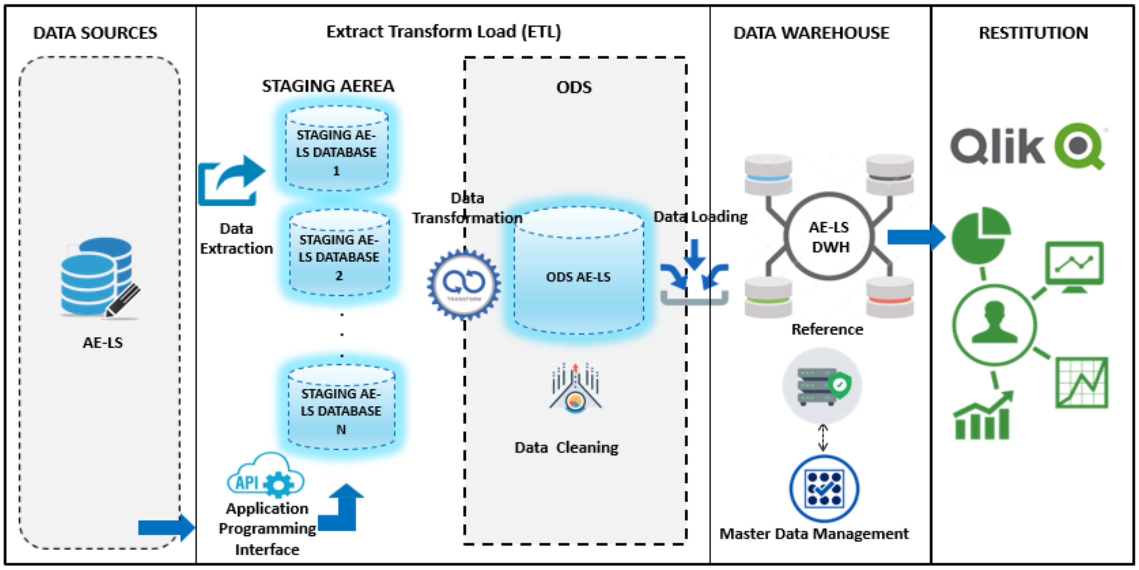


Fig. 3. Suggested Learning Analytics System.

III. Suggested Architecture

A. ETL Process Layer

The ETL architecture is organized into two areas: Staging Area and Operational Data Store (ODS). Each area consists of a separate database to store data provided by AE-LS.

1) Data source extraction: The proposed system uses multiple data sources of AE-LS such as Open Educational Data sets, E-learning system, LMS, AHS, and DAHS. To extract data from AE-LS database, the analytic system uses the Application Programming Interface (API) function, especially the Data Manipulation API and the Logging Data API. The data manipulation API is used to extract the learner data from the relational databases of AE-LS and import it in the Staging Area. While, Logging, Data API is used to record the data from AE-LS log files.

2) Staging area: Staging Area is created for every AE-LS data source that is activated. It has the same structure as their respective data source. It is also flagged with key fields for the request ID, the data package number, and the data record number. Staging Area is a copy of the data source that centralize data from heterogeneous sources to facilitate analysis operations. Its contents can be emptied after each load of the data warehouse.

In the staging area, the learner data is stored without cleaning or controls. Once the learner data is stored in the staging area, we can consolidate data in Operational Data Store (ODS) layer.

3) ODS: The integration of the data into the ODS implies a purge of redundant information and incoherent learner data. This stage is attempted to consolidate data, remove duplicate data, clean data, and change their format to the adequate format. All reconciliations and aggregations are stored in the ODS layer to obtain the desired structure for learner's behavior analysis. The result of transformations is loaded in the Data Warehouse.

B. Data Warehousing Layer

In order to build an analytic model for AE-LS, we propose a model of DWH implemented in PostgreSQL database management system (DBMS). The proposed model covers all data and analysis issues covered by adaptive e-learning systems. The DWH is organized in star schema and hosted on a PostgreSQL server. It collects data from a variety of diverse and heterogeneous sources of AE-LS, which the primary purpose is to support analysis and facilitate the decision-making process. This DWH store a large amount of historical data and enable faster and complex queries across data and its derivatives or even in-memory. The database is used to store current transactions and allows quick access to specific learner's transactions, including indexing technologies. It is usually standardized, which means that there is a single copy of each data.

The Data flow of this approach defines which objects are needed at design time and which processes are needed at runtime. These objects and processes are needed to transfer data from the AE-LS source to the ETL and DWH, to cleanse, consolidate, and integrate the learner' data, so that it can be used for analysis, reporting and planning. The individual requirements of these processes are supported by numerous options for designing the data flow. Fig. 4 shows the data flow architecture for the proposed learning analytics system.

To ensure the quality of AE-LS data in a data warehouse, we opt to use a Master Data Management (MDM) solution. The MDM solution helps to increase their overall efficiency by better use of the analytics system. The implementation of the MDM solution in the DWH improves the process of data and allow a quicker and simpler management of the new activities of learners. It gives the possibility to eliminate unnecessary data and processes, make data reusable, improve overall data quality, and define reference tables.

C. Restitution Layer

The restitution layer receives the result data of the star schema from PostgreSQL DBMS and visualizes it with Qlik Sense tool. Qlik Sense Allows to explore heterogeneous educational data, it makes it easy to choose data with a single click and offers a quick grip and setting. It is faster, more intuitive and less bulky than multidimensional OLAP solutions.

This solution is an end-to-end application, which contains the analysis objects, their setting, and the data to be analyzed. Qlik Sense uses high-performance technology that can process millions of recordings in memory; when constituting the data model, the AQL (Associative Query Language) technology proceeds automatically by homonymy, and indexes existing logical links in the data. This allows an exemption from the definition of relational joins when developing and using dashboards. Fig. 5 illustrates the Qlik Sense architecture.

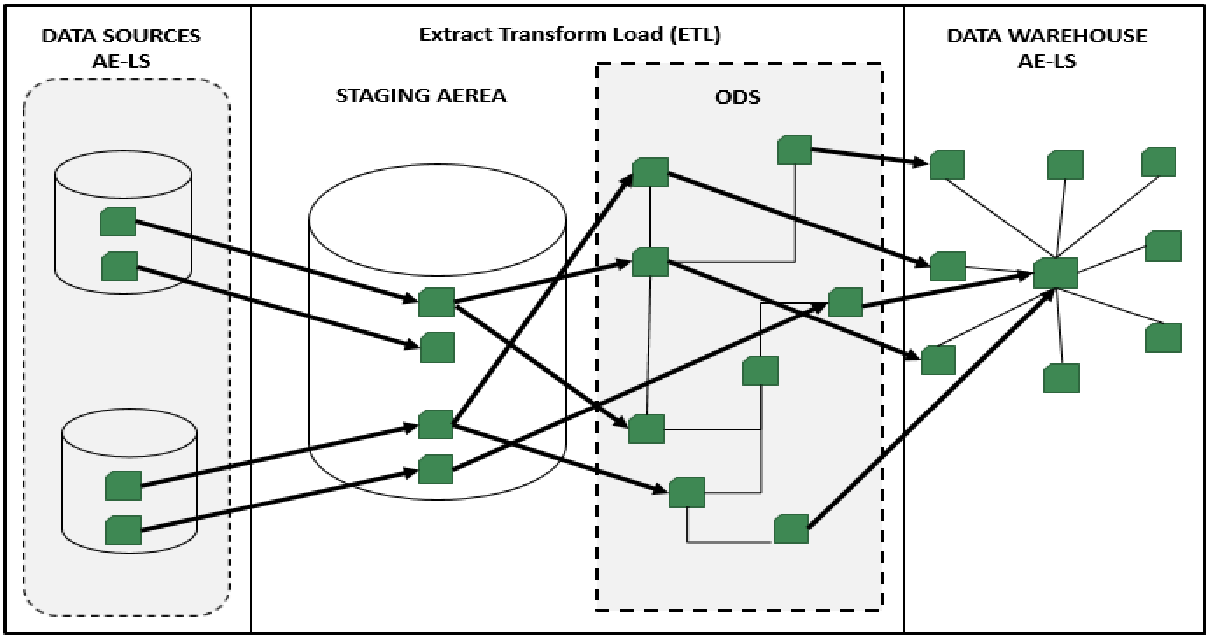


Fig. 4. Data Flow Architecture for the Proposed Analytics System.

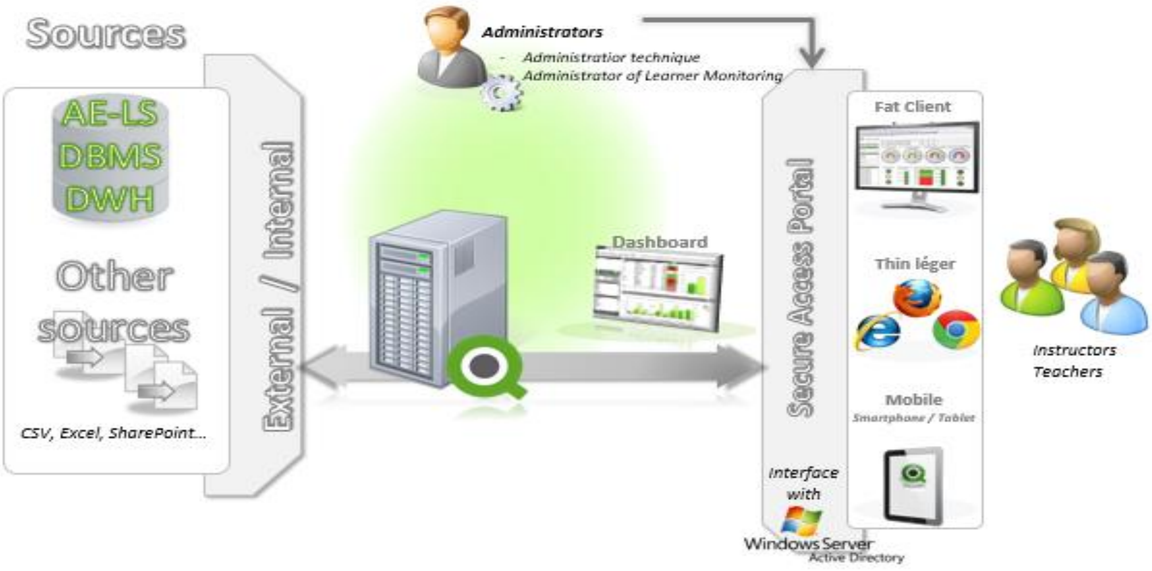


Fig. 5. Qlik Sense Architecture.

IV. Case Studiy of an Anonymous AE-LS Data Set

In the case study, we use the anonymised Open University Learning Analytics Dataset (OULAD) [31]. It contains data about, learners, courses and their interactions with AE-LS. We extract data from the open dataset of AE-LS anonymously according to the ethical and privacy of the system. The AE-LS dataset contains data from years 2013 and 2014.

The data components of the system architecture include the data sources, which correspond to the anonymized Open University Learning Analytics Dataset. From this dataset, data are extracted, transformed and loaded into the data warehouse, where data is stored in a suitable format.

A. Talend Architecture Job

Talend is the ETL solution, which will be used. It not only guaranties the highest level of performance but also the most cost-effective solution available in BI. It is an open source desktop application offers intuitive drag-and-drop tools, smart guides and automated processing features. It allows users to explore, clean, enrich and combine data from different sources. It allows users to simplify and speed up the often time-consuming and laborious process of data preparation, management, manipulation and data analysis tasks using spreadsheets. Fig. 6 presents the ETL Job using Talend tool.

In the Talend architecture jobs, we extract data from flat files of the anonymous dataset using File Input Delimited job and store data in Staging area using separate staging area databases. There are neither restrictions nor integrity key is used, the data are treated as if the tables were flat files. The second step is transforming the data in order to store them in ODS. Table II presents an example of transformations used in this approach.

TABLE. II. Example of Transformations used in the Suggested Approach

CSV Sources	Transformations	Components
StudentRegistrat-ion StudentInfo	Join data together from Learner and learner_Info sources.	tJoin
Courses Assessments	Remove duplicate data.	tUniqRows
Assessments	Select only certain columns.	tFilterColumns
StudentVLE	Extract Day, Month, and Year from date.	tMap: Day(date) /Month(date) /Year(date)
StudentInfo	Convert data into capital letters.	tMap: StringHandling.Upcase(region)
Courses	Delete spaces on the right.	tMap : StringHandling.TRIM(Label_courses)
VLE	Replace with '0' if null.	tMap: Relational.ISNULL(sum_click) ? "0" : sum_click
StudentInfo StudentRegistrat-ion Courses VLE StudentVLE StudentAssessm-ent Assessments	Construct metrics and statistics of the learner's activities based on values and calculations.	tAggregatorRows

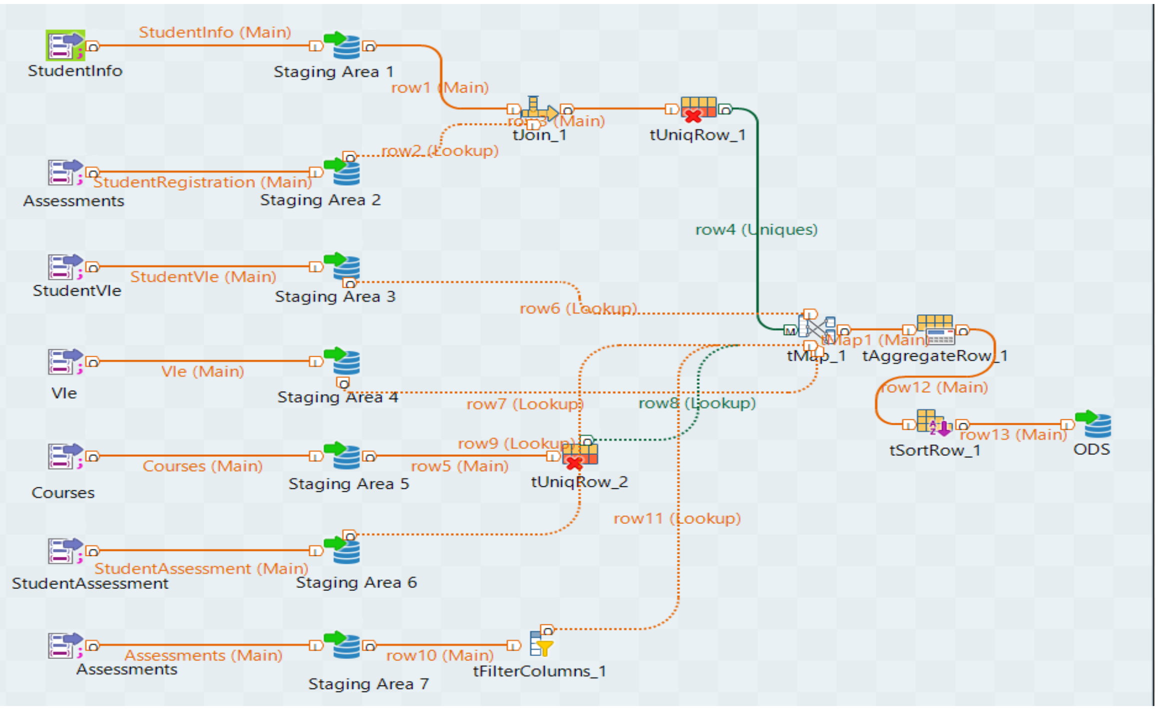


Fig. 6. ETL Job of the Suggested Approach using Talend Open Studio.

B. Indicators Analysis and Measures Identification

The ODS layer of ETL process contains the most important transformations' data, which be stored in ODS. This layer is intended to contain indicators, measures, and data of the finer granularity. In this approach, we mean by indicator, the aggregation of data, generally quantitative, called measures. Thus, we searched for indicators to measure learner's activities. The particularity of these indicators is that are applied to study learner's activities such as:

- Number of connections;
- Time of connection by day;
- Last access to the content;
- Number of visits courses;
- Time spend viewing learning content;
- Preferred content learning resources;
- Preferred learning resource;
- The manner which learners use the collaborative tools;
- Preferred learning activities of learners, etc.

These indicators and measures will be stored in the DWH, especially in the Fact table according to their dimensions.

C. Dimensions Identification

In the proposal, the dimension is an element constituting the context of the indicator. It is a structure that categorizes measures in order to enable teachers to answer pertinent questions about learner's activities.

1) Assessments_DIM: Assessments_DIM contains the results of learner evaluations. If the learner does not submit the assessment, no results are recorded. In addition, submissions for the final exam are missing if the assessment result is not stored in the system.

This dimension includes the identification number and the name of the assessment, a status flag indicating that the assessment result has been transferred from a previous module and the learner's score on this assessment.

2) Time_DIM: When a dimension is updated, the ETL keeps the memory of the old values in order to the old facts remain linked to the old values. To manage the history of dimensions' values in the data warehouse, we associate dates with other dimensions and facts in order to know what dimension value is valid for what fact, using the Time_DIM dimension. It is structured as a hierarchy of minutes, hours, days, months, and years of activities. The suggested model of DWH, refresh data periodically.

3) Learners_DIM: Learners_DIM contains learners' information, such as learner ID, the age band, gender, level of education, disability, etc.

4) Modules_DIM: Module_DIM contains the information about each module such as name, type, and levels of education in the module.

5) Courses_DIM: This dimension contains the list of all available courses to learners and their identification numbers.

*6) Content_Presentation_DIM:*Presentation_content_DIM contains information about the available content presentation materials in the AE-LS. Typically, these presentation materials are html pages, text, power point, audio, video, etc. Learners have access to these contents presentation online and their interactions with the materials are recorded. It includes the content presentation ID, the type of activity associated with the content presentation materials, and the learning resources.

In this approach, we use Talend MDM to manage the reference data. The reference data is data used to categorize other data within applications and databases. It refers to this data as a look-up, a code or a value. This table is characterized by code and value such as the two reference tables of the proposed DWH: Lanquage_DIM and Geography_DIM.

7) Lanquage_DIM: Language_DIM is a reference table, which refers to categorize language who learner use, say or write when they are trying to communicate spontaneously with learners or with e-learning system.

8) Geography_DIM: Geography_DIM is a reference table, which refers to identify the geographic region, where the learner lived while taking the module-presentation.

D. Data Warehouse Modelling

In this section, we suggest a data warehouse modelling based on identified indicators, measures and dimensions previously presented. The data warehouse modelling consists of one fact table and eight dimensions previously mentioned: Learner_Activities_FACT, Time_DIM, Leaner_DIM, Modules_DIM, Courses_DIM, Assessments_DIM, Content_Presentation_DIM, Language_DIM, and Geography_DIM.

1) Learner_Activities_FACT: The fact table, Learner_Activities_FACT, accumulates in the data warehouse; there is never any deletion or update (monotonous growth) to manage the history and traceability of data. It is linked to each dimension table with its foreign key. The fact table contains the learner activity ID, some attributes of learner's activities, and the foreign keys of these dimensions. The main measure of the fact table are : number of visits courses, time spend viewing learning content , last access to the AE-LS, preferred learning resource, preferred language, preferred content presentation, preferred learning activities, the number times which the learner has attempted this module, the total number of credits for modules which the learner is currently studying, learner's final result, the Index of Multiple Depravation Band of the place where the learner lived during the module, and the number of the learner's interactions with the materials in the AE-LS, etc. Fig. 7 presents the suggested data warehouse modelling for deploying analytics in AE-LS.

E. Results

Fig. 8 illustrates an example of dashboards using the suggested analytic system and the anonymised dataset of AE-

LS. The information used in this reports help teachers to have a visibility of the learner's activities and know their preferences. Thus, it allows to improve the presentation of learning resources and the performance of learners taking into consideration the band of learner's ages.

In the year 2014, 46.4% of learners prefer to participate in forums and virtual class, which allows them to interact directly with teachers and other learners. Around 23.1% of learners stay on the home page, which shows the obligation to improve the presentation of the home page to attract learners' attention and motivates them to attend courses and pass quizzes. In contrary, 3% of learners access to the content learning resources. For this reason, the e-learning system should use the interactive learning resources such as interactive videos, speeches and graphical simulations all over for learners who suffer from visual and auditory limitations that meet to their needs.

F. Limitations

This paper contributes architecture of analytics systems of AE-LS to monitor learner's activities. This approach uses a model of data warehousing as a result of the process of the extraction, transformation, and loading data, which be explored in the restitution using Qlik Sense tool. The architecture is applicable for any e-learning data sources where the data types can be partitioned into numeric, date/time, and textual data types.

However, this approach is limited in the capability of handling unstructured data. Structured data has a fixed schema so that to fit it a relational model can be designed. On the other hand unstructured and semi structured data do not have a fixed schema. Nowadays in e-learning system, these types of data are becoming common. These data also hide prudent information. Therefore, the decision-making in the learning process would not be effective as these unstructured data are not suitable.

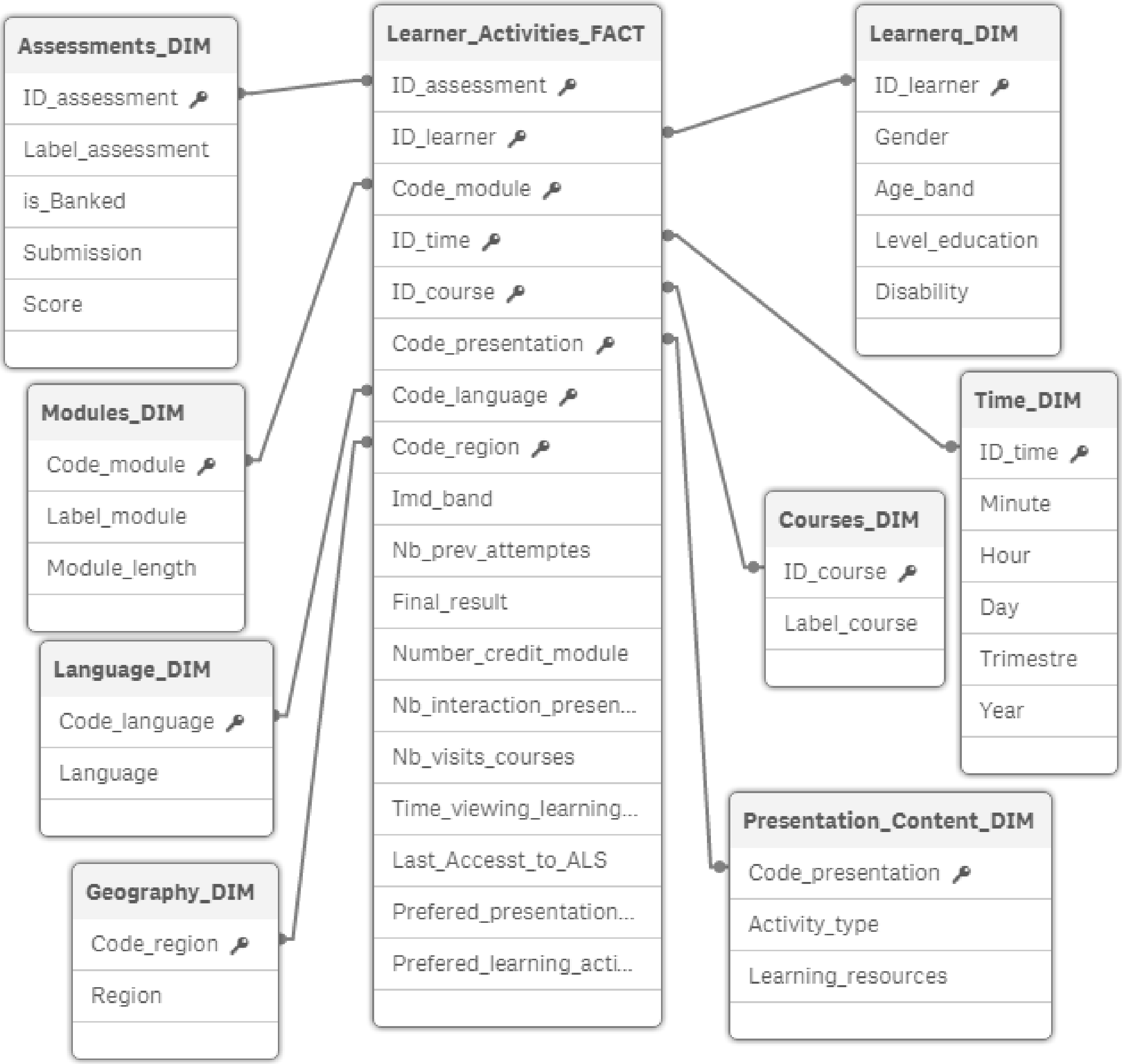


Fig. 7. Suggested Data Warehouse Modelling of Learning Analytics in AE-LS.

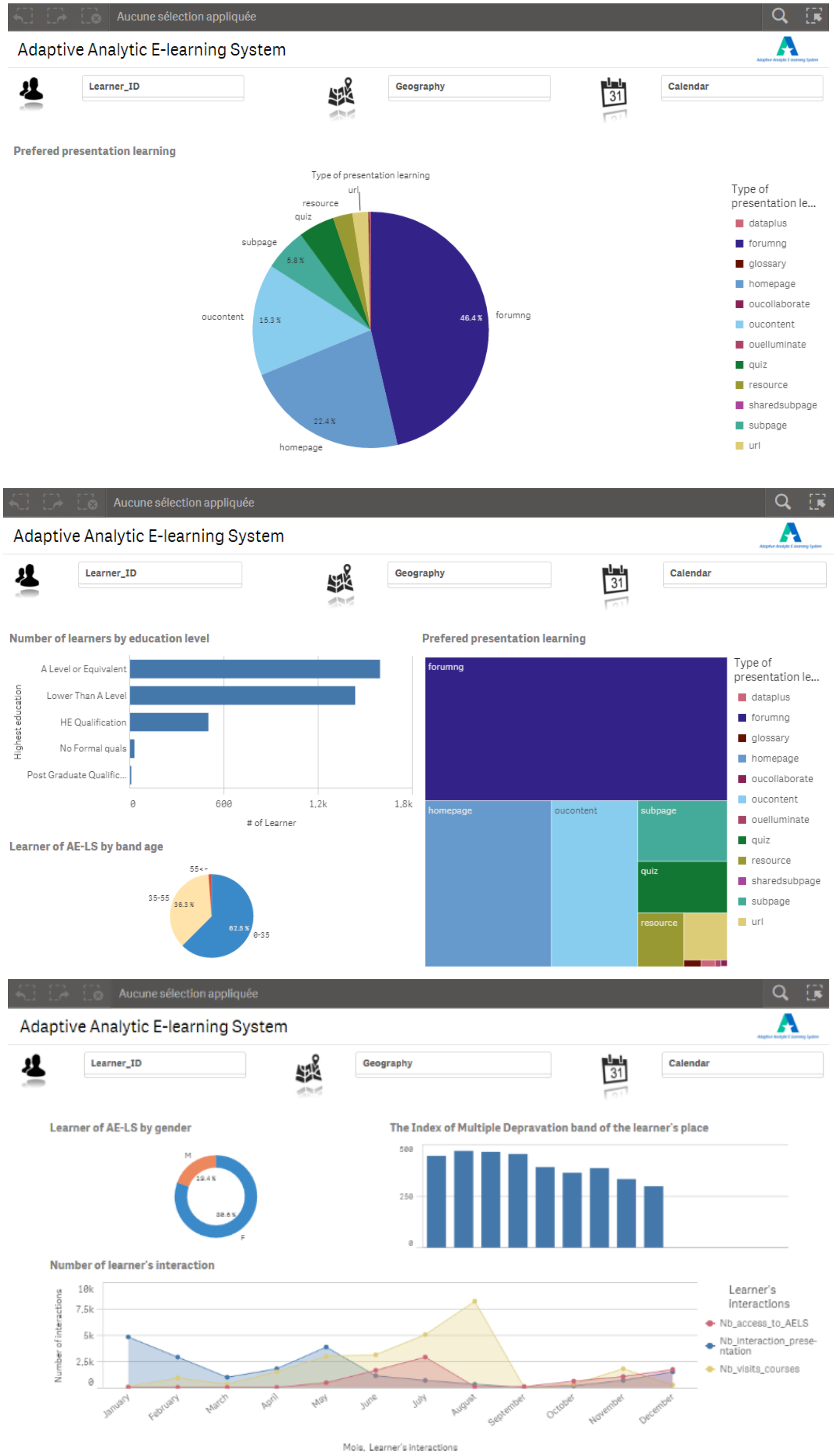


Fig. 8. Example of Dashboards using the Suggested Analytics System.

V. Conclusion

In this paper, we present a new approach to monitor and perform the learner's behavior. The suggested framework tries to solve the lack of reports in an analytics system according to AE-LS. This approach provides teachers and administrators with detailed reports to assess, evaluate and perform learners' activities. It allows instructors to visualize, analyze and discover learners' behavior in order to improve the quality of teaching and take the suitable decision.

This study shows that learner spends little time on courses and tutorials and prefer to use collaborative activities, such as virtual classes and forums, rather than looking at learning resources. In future work, we will use a predictive model of data mining to help instructors to have an idea of future learner's behavior, and propose to the learner an adaptive presentation content.

References

[1] KLOCK, Ana Carolina Tomé, GASPARINI, Isabela, PIMENTA, Marcelo Soares, et al. Adaptive hypermedia systems. In : Advanced Methodologies and Technologies in Media and Communications. IGI Global, 2019. p. 217-228.

[2] RAJI, Mohammad, DUGGAN, John, DECOTES, Blaise, et al. Visual progression analysis of student records data. In : 2017 IEEE Visualization in Data Science (VDS). IEEE, 2017. p. 31-38.

[3] EL JANATI, Salma et MAACH, Abdelilah. Towards a new adaptive E-learning framework for adapting content to presentation. In : 2017 Intelligent Systems and Computer Vision (ISCV). IEEE, 2017. p. 1-7.

[4] SFENRIANTO, Sfenrianto, HARTARTO, Yustinus B., AKBAR, Habibullah, et al. An Adaptive Learning System based on Knowledge

[5] Level for English Learning. International Journal of Emerging Technologies in Learning (iJET), 2018, vol. 13, no 12, p. 191-200.

[6] WANG, Feng Hsu. An exploration of online behaviour engagement and achievement in flipped classroom supported by learning management system. Computers & Education, 2017, vol. 114, p. 79-91.

[7] HILLES, Mohanad M. et NASER, Samy S. Abu. Knowledge-based Intelligent Tutoring System for Teaching Mongo Database. 2017.

[8] KLOCK, Ana Carolina Tomé, GASPARINI, Isabela, PIMENTA, Marcelo Soares, et al. Adaptive hypermedia systems. In : Advanced Methodologies and Technologies in Media and Communications. IGI Global, 2019. p. 217-228.

[9] PAPANIKOLAOU, Kyparisia A., GRIGORIADOU, Maria, KORNILAKIS, Harry, et al. Personalizing the Interaction in a Web-based Educational Hypermedia System: the case of INSPIRE. User modeling and user-adapted interaction, 2003, vol. 13, no 3, p. 213-267.

[10] KNUTOV, Evgeny, DE BRA, Paul, et PECHENIZKIY, Mykola. AH 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. New Review of Hypermedia and Multimedia, 2009, vol. 15, no 1, p. 5-38.

[11] EL JANATI, Salma, MAACH, Abdelilah, et EL GHANAMI, Driss. SMART Education Framework for Adaptation Content Presentation. Procedia Computer Science, 2018, vol. 127, p. 436-443.

[12] BRUSILOVSKY, Peter. Adaptive navigation support: From adaptive hypermedia to the adaptive web and beyond. PsychNology Journal, 2004, vol. 2, no 1, p. 7-23.

[13] MÖDRITSCHER, Felix, BARRIOS, Victor Manuel García, et GÜTL, Christian. Enhancement of SCORM to support adaptive E-Learning within the Scope of the Research Project AdeLE. In : E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education. Association for the Advancement of Computing in Education (AACE), 2004. p. 2499-2505.

[14] SCLATER, Niall. Learning analytics explained. Routledge, 2017.

[15] ROMERO, Cristobal et VENTURA, Sebastian. Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 2007, vol. 33, no 1, p. 135-146.

[16] Wixom, B. H., Watson, H. J., and Werner, T. 2011. Developing an enterprise business intelligence capability. MIS Quart. Exec. 10, 2.

[17] Turban, E., Sharda, R., Aronson, J. E., and King, D. 2008. Business Intelligence: A Managerial Approach. Pearson Prentice Hall.

[18] W. H. Inmon, Building the Data Warehouse. Chichester: Willey & Son, 2002.

[19] R. Kimball and J. Caserta, The Data Warehouse ETL Toolkit. Chichester: John Wiley & Sons, 2002.

[20] R. Mazza and V. Dimitrova, "CourseVis: A graphical student monitoring tool for supporting instructors in web-based distance courses," International Journal of Human-Computer Studies, vol. 65, no. 2, pp. 125–139, 2007.

[21] LIU, Bing. Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media, 2007.

[22] HAN, Jiawei, KAMBER, Micheline, et PEI, Jian. Data mining: concepts and techniques. 2001. San Francisco: Morgan Kauffman, 2006.

[23] GARFIELD, Joan et BEN-ZVI, Dani. How students learn statistics revisited: A current review of research on teaching and learning statistics. International statistical review, 2007, vol. 75, no 3, p. 372-396.

[24] MAZZA, Riccardo. Introduction to information visualization. Springer Science & Business Media, 2009.

[25] JAN, Shazia K., VLACHOPOULOS, Panos, et PARSELL, Mitch. Social Network Analysis and Learning Communities in Higher Education Online Learning: A Systematic Literature Review. Online Learning, 2019, vol. 23, no 1.

[26] M. E. Zorrilla and E. Alvarez, "MATEP: Monitoring and Analysis Tool for E-learning Platforms," in 8th IEEE International Conference on Advanced Learning Technologies, Santander, Spain, 2008, pp. 611-613.

[27] FRANCE, Laure, HERAUD, J.-M., MARTY, J.-C., et al. Monitoring virtual classroom: Visualization techniques to observe student activities in an e-learning system. In : Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06). IEEE, 2006. p. 716-720.

[28] AVOURIS, Nikolaos, KOMIS, Vassilis, FIOTAKIS, Georgios, et al. Logging of fingertip actions is not enough for analysis of learning activities. In : 12th International Conference on Artificial Intelligence in Education, AIED 05 Workshop 1: Usage analysis in learning systems. 2005. p. 1-8.

[29] MAZZA, Riccardo et MILANI, Christian. Gismo: a graphical interactive student monitoring tool for course management systems. In : International Conference on Technology Enhanced Learning, Milan. 2004. p. 1-8.

[30] J. Mostow et al., "An Educational Data Mining Tool to Browse TutorStudent Interactions: Time Will Tell!," in Workshop on educational data mining, 2005, pp. 15–22.

[31] MERCERON, Agathe et YACEF, Kalina. Tada-ed for educational data mining. Interactive multimedia electronic journal of computer-enhanced learning, 2005, vol. 7, no 1, p. 267-287.

[32] KUZILEK, Jakub, HLOSTA, Martin, et ZDRAHAL, Zdenek. Open university learning analytics dataset. Scientific data, 2017, vol. 4, p. 170171.

Machine Learning Approaches for Predicting the Severity Level of Software Bug Reports in Closed Source Projects

Aladdin Baarah[1], Ahmad Aloqaily[2], Zaher Salah[3], Mannam Zamzeer[4], Mohammad Sallam[5]
Department of Software Engineering, Hashemite University, Zarqa, Jordan[1, 5]
Department of Computer Science and its Applications, Hashemite University, Zarqa, Jordan[2]
Department of Computer Information Systems, Hashemite University, Zarqa, Jordan[3]
Department of Information Technology, the University of Jordan, Amman, Jordan[4]

***Abstract*—In Software Development Life Cycle, fixing defect bugs is one of the essential activities of the software maintenance phase. Bug severity indicates how major or minor the bug impacts on the execution of the system and how rapidly the developer should fix it. Triaging a vast amount of new bugs submitted to the software bug repositories is a cumbersome and time-consuming process. Manual triage might lead to a mistake in assigning the appropriate severity level for each bug. As a consequence, a delay for fixing severe software bugs will take place. However, the whole process of assigning the severity level for bug reports should be automated. In this paper, we aim to build prediction models that will be utilized to determine the class of the severity (severe or non-severe) of the reported bug. To validate our approach, we have constructed a dataset from historical bug reports stored in JIRA bug tracking system. These bug reports are related to different closed-source projects developed by INTIX Company located in Amman, Jordan. We compare eight popular machine learning algorithms, namely Naive Bayes, Naive Bayes Multinomial, Support Vector Machine, Decision Tree (J48), Random Forest, Logistic Model Trees, Decision Rules (JRip) and K-Nearest Neighbor in terms of accuracy, F-measure and Area Under the Curve (AUC). According to the experimental results, a Decision Tree algorithm called Logistic Model Trees achieved better performance compared to other machine learning algorithms in terms of Accuracy, AUC and F-measure with values of 86.31, 0.90 and 0.91, respectively.**

Keywords*—*Software engineering; software maintenance; bug tracking system; bug severity; data mining; machine learning; severity prediction; closed-source projects

I. Introduction

Over the past years, the process of fixing bugs in the software maintenance phase has become a challenging task because the number of reported software bugs in large software systems (e.g., open-source projects) is growing massively [1]. For instance, the number of reported bugs submitted daily to Mozilla open-source system are 135 on average [2]. Managing a high volume of new bugs submitted daily by different reporters in these large open systems is a difficult task. Furthermore, this increases the amount of work done by the so-called bug triager, who assesses and analyzes these bugs within the bounds of time and resources to assign an appropriate severity level and suitable developer(s) who will fix those defected bugs [3].

Bug report systems are important software artifacts. They are leveraged for various tasks during software maintenance, such as assigning bugs to appropriate developers, assessing the severity and priority of bugs and detecting duplicate bugs. The quality of bug reports relies, to a large extent, on the information written in these reports [4]. Thus, if the data in bug reports are incomplete, unclear or inaccurate, the tasks as mentioned above will lead to unpredictable results.

Bug tracking systems such as Bugzilla [5] and JIRA [6] are utilized by open-source and closed-source projects during the software maintenance phase to collect, maintain, manage, and track issues related generally to bug reports (i.e. corrective maintenance) [7]. At present, developers, quality assurance testers, users and other team members are promoted to report and submit bugs they experience to bug repositories in a situation when the project behaves incorrectly and does not conform to the software requirements [8].

Bug severity “is the degree of impact that a defect has on the development or operation of a component or a system” [9]. Generally, bug reports are categorized according to their severity. High severity reports exemplify major (i.e. critical or fatal) errors and have a high impact on the functionality of the system, as well, they are given more top priority and should be resolved rapidly. While low severity reports exemplify trivial errors and minor problems that do not affect the execution of the system [10].

Initially, when a new bug is created, the reporter (e.g., developer) is required to fill in the fields of the bug report form such as a one-line summary for a specific project. As well, the reporter is required to estimate the severity level field (e.g., high, medium and low) of the observed bug according to their competence and knowledge. In practice, if the reporter is incapable of assessing the severity of the bug, they will assign a default value for the severity field and thus makes the triaging process more difficult [11]. One potential explanation for assigning a default value for the severity is that the reporters are unable to distinguish between different severity levels, or they have lack of knowledge and experience in assigning the value of the severity level [12]. Another potential explanation is that the reporters do not take into consideration the severity estimation at all when they submit a report and leave the default value unchanged [10].

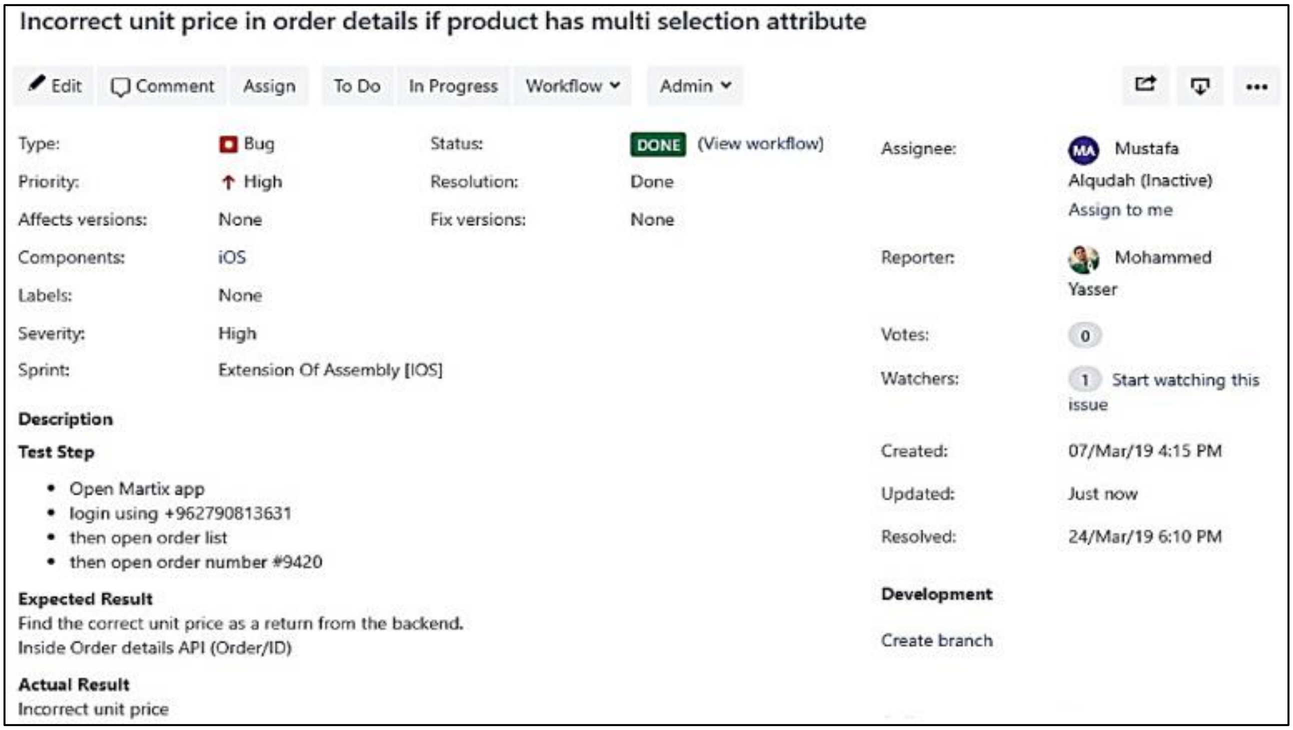
Even though there are rules on how to set the severity label of the encountered bugs, specifying proper severity level is mostly based on the expertise of who reports the bug. The developer has to determine the severity of these software bugs manually to prioritize which bug report needs more attention than others in terms of fixing. In the case of assigning an inaccurate value for the severity, this will result in postponement in fixing severe bugs, and this will expand the time to resolve these type of bugs [13]. However, a further validation step is required after the bug has been submitted where the bug triager has to confirm the validity of the severity value, whether it is adequately assigned or not. This process is called "severity identification" [14]. It is a tiresome and time-consuming process and it increases the volume of work carried out by the triager especially when there is a massive number of bugs submitted daily to the bug tracking systems.

Fig. 1 shows an example of the JIRA bug report used by INTIX Company. This bug report is reported by "Mohammad Yasser" and assigned to the developer "Mustafa Alqudah". The one-line summary "Incorrect unit price in order details if a product has multi-selection attribute" is shown in the top left of the figure and the severity level assigned for this bug is high.

To overcome the mentioned problems, there is a necessity to automate the whole process of assigning the severity level of newly reported bugs to replace the manual job. One reason for that is to decrease the amount of work done by the triager. Another reason is to improve the accuracy of severity identification.

In this paper, we compare different machine learning algorithms applied on the textual description of the bug reports (i.e. one-line summary field). Unlike most of the research works reported in the literature, we applied the proposed methodology to a private dataset related to bug reports. These bugs are associated with closed-source projects developed by INTIX Company located in Amman, Jordan. The dataset is extracted from JIRA bug tracking system used by the company.

The structure of this paper is organized as follow. In Section II, we describe the related works conducted by other researchers. Then, in Section III, we demonstrate our research methodology in detail. After that, in Section IV, we discuss the experimental results. Finally, we summarize our approach and point out future work in the conclusion section.



Incorrect unit price in order details if product has multi selection attribute

Type: Bug
Priority: High
Affects versions: None
Components: iOS
Labels: None
Severity: High
Sprint: Extension Of Assembly [IOS]

Status: DONE
Resolution: Done
Fix versions: None

Assignee: Mustafa Alqudah (Inactive)
Reporter: Mohammed Yasser
Votes: 0
Watchers: 1
Created: 07/Mar/19 4:15 PM
Updated: Just now
Resolved: 24/Mar/19 6:10 PM

Development

Description

Test Step

- Open Martix app
- login using +962790813631
- then open order list
- then open order number #9420

Expected Result

Find the correct unit price as a return from the backend.
Inside Order details API (Order/ID)

Actual Result

Incorrect unit price

Fig. 1. Example of JIRA Bug report.

II. Related Works

The work described in this paper is concerned mainly with bug reports, particularly in the area of bug severity. This section presents the recent literature review regarding bug severity prediction.

One of the first studies to predict the severity label of bug reports was performed by [15]. A rule-based learning technique was utilized to build a new tool called SEVERIS. SEVERIS is based on text mining and machine learning techniques applied on the unstructured data of the bug report (i.e. summary, description). They applied their automated prediction model on NASA's Project and Issue Tracking System (PITS). The authors argued that SEVERIS, with a slight adjustment, can be applied to other open-source repositories such as Bugzilla.

An extension of [15] work was performed by [11] to investigate whether text mining techniques applied in the textual information of a bug report is an adequate approach to predict the severity of a newly reported bug automatically. They used Naïve Bayes technique for their prediction model and they applied their model on datasets extracted from three popular open-source bug repositories, including Mozilla, Eclipse and GNOME. The experimental results showed that prediction accuracy was varied between 0.65–0.75 for Mozilla and Eclipse dataset. Regarding GNOME dataset, the prediction accuracy was varied between 0.70–0.85.

Later, a follow-up study was conducted by [10]. The authors compared four familiar classification techniques (Naïve Bayes, Naïve Bayes Multinomial (NBM), K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM)) to figure out which technique was the most suitable for the severity prediction model so that it can classify the newly reported bug into severe and non-severe. For evaluation purposes, they construct twelve different datasets from Eclipse and GNOME open-source projects. According to their experimental results, among the four candidate classifiers, NBM had the best results in terms of accuracy.

A comparable study to [15] was conducted by [16]. The authors conducted different experiments on four nameless datasets of NASA's PITS using three methods: a Regression method called Multi-Nomial Multivariate Logistic Regression (MMLR), Multi-Layer Perception (MLP) and Decision Tree. The prediction models were fed with different top-k terms extracted from the training datasets using Information Gain (IG) feature selection. The authors concluded that the Decision Tree performed better than MMLR and MLP.

A further study for predicting fine-grained severity level of a bug report was conducted by [17]. The authors built a model to classify the severity of a new bug reported to the bug repository. A dataset of 163 bug reports was built from Eclipse and Mozilla projects. Six machine learning approaches were adopted, namely, Naïve Bayes, RBF Networks, Functional Trees, Random Trees, Random Forests and AdaBoost. In accordance with their results, AdaBoost with base classifiers, as mentioned above, showed an improvement of up to 4.9% in terms of accuracy.

Other studies, for example [18, 19], examined the impact of utilizing other attributes of bug reports, besides the unstructured text, to enhance the bug severity prediction. In the study conducted by [18], the authors used NBM as a classification technique and applied their proposed approach on two distinct datasets originating from two open-source projects, namely Mozilla and Eclipse. Their empirical study revealed that there was an improvement in the prediction accuracy and outperformed the work of [11]. Similarly, Yang et al. [19] adopted NBM classifier for severity prediction. According to their evaluation, the attributes (quality indicators) of bug reports from Eclipse dataset (i.e. stack traces, attachments) exploited in their study showed an improvement in the accuracy of their prediction model.

In the domain of cross-project severity prediction, Singh et al. [20] conducted a study based on the summary description of the bug report. Different prediction models were built using different classifiers, namely SVM, K-NN and Naïve Bayes. A combination of seven datasets from seven Eclipse projects was constructed to build 63 training set candidates. The experimental results revealed that K-NN achieved better results than SVM and Naïve Bayes when using a combination of several training datasets rather than a single training dataset.

Many research works, as described in [21-25] employed different feature selection methods to minimize the number of informative features set and to enhance the accuracy of the classifier. All the works mentioned above paid particular attention to the effectiveness of feature selection techniques on the accuracy of classifying the severity of bug reports. Bi-gram, along with feature selection and text mining algorithms were proposed by [22] to enhance the accuracy of their prediction model. While in work described by [25], the authors examined three feature selection methods, namely, Information Gain (IG), Chi-Square (CHI) and Correlation Coefficient to extract the distinct features that depict the severity of the bug reports (severe or non-severe) from the summary field in the bug reports.

On the other hand, an ensemble feature selection technique was proposed by [21] through consolidating at most two feature selection methods. Various feature selection methods were used in their experiments, in particular, Term Frequency (TF), Document Frequency (DF), Mutual Information (MI), Statistical Dependency (SD), Information Gain (IG), Chi-Square (CHI) and Correlation Coefficient. In their prediction model, the authors employed Naïve Bayes Multinomial as the classifier.

Later, [26] proposed a new approach, an integration of topic modelling using LDA and similarity using KL-divergence to predict the severity of bug reports in cross projects. In their work, a total of 20,000 bug reports from 4 different open source projects (Eclipse, Mozilla, WireShark and Xamarin) were assembled to validate their proposed approach. The experimental results demonstrated that their model achieved better performance, in terms of accuracy than other four cutting-edge studies listed in their literature.

In more recent studies, emotional-based expressions written in the unstructured text fields of the bug reports were leveraged by [27, 28] to classify the severity label of bug report (i.e., critical, high, trivial and low). In the former approach [27], two stages related to emotional similarity technique were performed using Smoothed Unigram Model and KL-Divergence. In the latter approach [28], the authors exploited the notion of deep learning algorithm based on Neural Network along with emotional analysis. The experimental results of both studies above emphasized that employing emotion analysis significantly improved the performance of the prediction model.

In the context of bug report severity prediction, most of the reported research studies in the literature mainly used open-source projects because they are publicly accessible. While in the research work described in this paper, we are attempting to exploit a private bug report dataset related to closed-source projects developed by an existing local company located in Jordan.

III. Research Methodology

The process of predicting the severity label of reported bugs is shown in Fig. 2. It comprises of four main phases: dataset extraction, dataset pre-processing, feature selection and prediction. The following sections will explain each phase in detail.

A. Dataset Extraction

The bug reports dataset is extracted from the repository of JIRA bug tracking system related to closed-source projects developed by INTIX Company located in Amman, Jordan. We considered bug reports submitted to JIRA between May 2016 and March 2018.

The bug reports were classified into five levels: lowest, low, medium, high and highest. Although software engineers usually follow a specific guideline on how to assign severity of reported bugs, however, the categorization process seems to be evaluated imprecisely. In this research work, we treat lowest and low severity as non-severe, while high and highest severity as severe bugs. Furthermore, as proposed by [12], the authors recommended not to take the medium severity in the classification process. The medium class, as it is investigated, is the default option for reporting a bug and it seems that a large number of reporters report any confusing bug as a medium.

The datasets mainly consist of three features including bug ID, a short description (i.e., one-line summary) and the severity level of the bug. The description of each bug is represented as a short text (snippet). From the bug reports, the severity and the short description of the bug report were mainly used for the prediction process. Table I statistically summarizes the characteristics of the dataset used in the experiments described in this research work. As shown in Table I, too many words are repetitive in the dataset where the number of distinct words is significantly fewer than the total number of all words exists within the dataset (765 out of 9016). The table also shows the shortness property of the bug summary (i.e. short text) where the average number of words per bug summary was 7.75 words and thus this made the prediction process more challenging.

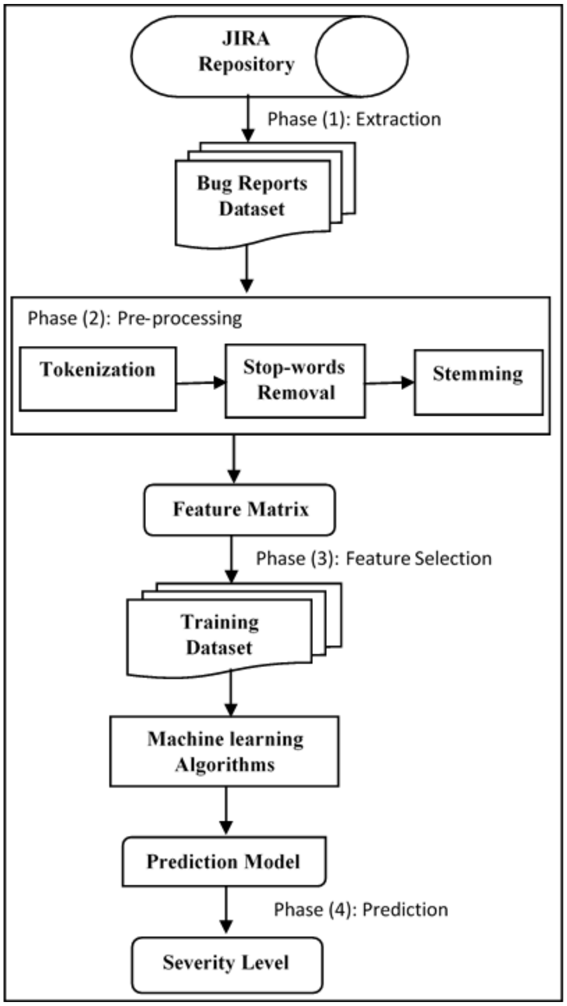


Fig. 2. Methodology for Predicting Bug Severity.

TABLE. I. Statistical Summary for the Bug Reports in JIRA Dataset

Min number of words per Bug Summary	1
Max number of words per Bug Summary	44
Avg. number of words per Bug Summary	7.75
StD. number of words per Bug Summary	3.73
Total Number of Bug Reports in the dataset	1164
Number of Severe Bugs	851
Number of Non-Severe Bugs	313
Number of all words in the dataset	9016
Number of distinct words in the dataset	765

B. Dataset Pre-Processing

To build a prediction model, machine learning algorithms require text data (i.e., one-line summary) of a bug report to be transformed into a feature of vector (Bag-of-Words). Therefore, different pre-processing steps must be applied to the dataset (text) to be converted into a vector of features where the features represent words in a utilized dataset. The typical pre-processing phase starts with Tokenization, Stop-words Removal and Stemming. These steps are explained as follow:

1) Tokenization: This step represents splitting a text data into a collection of words where each word corresponds to a single term. Removing white spaces, punctuations and converting all uppercase characters to lowercase are taken into consideration in this step.

2) Stop-words removals: In the domain of natural language processing, conjunctions, adverbs, prepositions and other constructive terms are used mainly to build a sentence. These terms do not carry out any semantic or statistical information to distinguish the text. Terms like "an", "on", "there" and many others are called stop-words and are not crucial for bug severity prediction. Therefore, all stop-words are removed based on a pre-defined list of stop-words.

3) Stemming: Stemming is one of the main steps in the domain of text mining and natural language processing. The purpose of carrying out this step is to replace all terms that have a common stem (root word) and the stem of a word is retained as a feature. For instance, the words "find", "finds", "found" and "finding" can be replaced with one word or their stem which is "find".

Finally, all the terms obtained after the three aforementioned pre-processing steps are called features or bag-of-words and are mainly used to build prediction models as described later in this section.

C. Feature Selection

The size of the feature set that can be generated after applying pre-processing steps is still enormous and consequently is not suitable for machine learning algorithms. Feature selection, in text mining, is a crucial step to select the most discriminative features (terms/words) from a list of features using appropriate feature selection methods.

There are many available feature selection methods in the literature such as Chi-square, Information Gain (IG), Term Frequency, Document Frequency, and Mutual Information. In fact, there is no much difference in utilizing different feature selection methods as we decide to select sets of features with different sizes [29, 30]. For experimental purpose, IG feature selection method is applied and features obtained after pre-processing are ranked and the top 'N' scoring features are selected based on ranking results to build the prediction models of different datasets.

IG feature selection is used to measure the dependency between features and the class label [31]. It measures the informativeness of a feature gained regarding the class label and is defined as follows:

$$IG(f_i, c_j) = H(f_i) - H(f_i|c_j)$$

where f_i is the feature (term) i and c_j is the class label j, $H(f_i\ )$ the entropy of term f_i, and $H(f_i|c_i)$ is the entropy of f_i after observing class label c_j. The entropy $H(f_i)$ is defined as

$$Entropy = -\sum_i p_i \ \log_2 p_i$$

Entropy is a common way to measure the degree of randomness or impurity of a variable and comes from information theory domain. The higher the entropy of a variable the more the information it holds about the class.

Finally, the bug reports are represented as a feature matrix where each row represents a bug report consisting of n selected features (terms). Each term is weighted using the Term Frequency Inverse Document Frequency (TF-IDF) approach. Term frequency is calculated by multiplying term frequency with inverse document frequency and is given as:

$$\text{TF-IDF} = n_w^d \log_2\left(\frac{N}{N_w}\right)$$

Where n_w^d the frequency of word w in document d, N is the number of document and N_w are documents containing word w. The *TF-IDF* value indicates that words (terms) which occur frequently in a specific document are more significant than other terms in the same document. Lastly, the features are normalized.

D. Machine Learning Algorithms

In this section, different machine learning algorithms for predicting the severity level of the bugs are described and utilized in this study. These algorithms are mainly concerned to deal with unstructured data such as text [32]. In this study, machine learning algorithms namely: Bayesian algorithm, Support vector machine, decision tree and rule-based algorithms are used to classify severity levels of bug reports. These algorithms are described in the following subsections:

1) Bayesian algorithms: Bayesian is a probability-based classification algorithm that is based on Bayesian theorem [33]. It describes the probability of an independent variable based on prior knowledge of the dependent variable(s). In this research study, two different variations of the Bayesian algorithm were used namely, Naïve Bayes classifier (NB) and Naïve Bayes Multinomial (NBM).

The Naïve Bayes classifier estimates the class conditional probability of class label with the assumption that all attributes are independent [34]. The NB has been widely used in the domain of text classification due to its simplicity and effectiveness. On the other hand, NBM is similar to the naïve Bayes classifier except that it considers a weight for each feature during probability calculations. The weight of each feature can be determined by specific distributions or as a parametric model. The parameters can be estimated based on the training data. For the case of NBM, the distribution of data is assumed as a multinomial model [35].

2) Support vector machine: The main idea underlying support vector machines (SVMs) is as follow: search the best maximal margin hyperplane separating two classes of a given training data. A margin is defined as the sum of the distances of the closest positive and negative correctly classified data points (support vectors) from the hyperplane while penalizing misclassified data points. In the case of linear classification problems, linear SVMs can be used to search for the hyperplane in the original space. On the other hand, for nonlinearly separable issues, the data are implicitly mapped to a higher dimensional space through a kernel function, where non-linear SVMs can be used to find a separating hyperplane.

For any given classification problem, if no hyperplane can separate the two classes, a soft margin approach can be used to control the sensitivity of outliers to allow a separating hyperplane. It determines a hyperplane that splits the cases as clearly as possible with a penalty for misclassified cases, while still maximizing the distance to the nearest support vectors. SVMs have been widely used in applications such as handwriting recognition and bioinformatics and showed superior performance [36-38]. Careful design and methodological approach must be taken in applying SVM algorithms including tuning of parameters. SVMs has also been applied in the domain of text mining as it shows satisfactory results. The high dimensionality of text data makes SVMs a useful algorithm to apply and also avoids the curse of dimensionality problem of text data [39]. An implementation of SVMs named SMO algorithm [40] has been applied to the dataset, which is an open-source implementation of SVMs.

3) Decision tree: The decision tree (DT) is a machine learning predictive approach used for classification and regression problems. It creates a prediction model by learning decision rules from data on a tree-like model. A DT model has a tree structure in which each node represents a test on a given variable and each branch represents the outcome of that test. Leaf nodes of the tree represent a class label (decision). The path from the root to leaf nodes represents a classification rule. Different DT based algorithms have been implemented including J48, random tree, random forest, Logistic model trees, and many others. Generally, decision tree classifiers have good accuracy. It is a typical inductive approach to learn knowledge from data.

In this study, three DT algorithms were applied. These algorithms include J48 [41], Radom Forest (RF) [42], and Logistic Model Trees (LMT) [43].

4) Rule-based algorithms: The Rule-Based Algorithm extract knowledge from data in the form of rules. Rules usually take the form of an if-then expression. The main feature of rule-based models is that the produced model is expressed in term of a set of rules rather than just one rule or model. In this study, an algorithm called (Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [44] is applied to the dataset. The RIPPER algorithm works through sequentially runs over the data in multiple passes. This algorithm repeatedly learns one rule at each pass until no data left. The JRip implementation of the RIPPER algorithm in WEKA is applied to the utilized dataset.

To conclude, the classification algorithms namely: NB, NBM, SVMs, J48, RF, LMT, JRip and KNN are mainly utilized to evaluate their performance on the studied dataset. The open-source Weka software [45] was used to run the experiments with different parameters setting and the best parameters setting are selected based on evaluation measures.

E. Experimental Settings

The validation approach used in our study to evaluate the performance of different classifiers is the k-fold cross-validation approach. This approach was used to generate a test set and avoid the over-fitting problem. The cross-validation approach performs independent tests without requiring separate test data samples and without reducing the data samples used to build prediction models. In the k-fold cross-validation process, the original data is classified into k groups, so that each group is used once as a validation set and the remaining data as a training set. In this study, the 10-fold cross-validation was used to evaluate the performance of different classifiers. Stratified based sampling was used to generate training and testing sets. In each case, a prediction model was trained using 9 folds of the data and tested on the remaining fold and the results were averaged.

F. Performance Evaluation

Once a classifier is trained successfully and a prediction model is generated, performance evaluation measures must be applied on a separate dataset called test set to evaluate the performance of the generated model. In our experiments, the performance of the generated prediction models is evaluated using various performance measures; these measures include accuracy, F-measure and Area Under the Curve (AUC). Each measure provides a different perspective of performance evaluation and a broader set of performance results to compare.

The accuracy measures how correctly a classifier predicts class labels. It is calculated as the percentage of true positive and true negative rates to the number of all instance. On the other hand, the F-measure considers both the precision and the recall to compute the performance of a classifier and is measured as the harmonic mean of precision and recall. In this case study, the *F1* measure is used, where recall and precision are equally weighted:

$$\text{F-measure } = 2 * \frac{precision * recall}{(precision + recall)}$$

The F-measure reach a value of 1 (perfect precision and recall) and the worst value is 0. The higher F-measure value indicates the higher quality performance of the classifier.

Furthermore, an alternative measure to evaluate the performance of generated models is the Receiver Operating Characteristic (ROC). This approach compares the true positive rate with a false positive rate as a drawn curve. The ROC measure is usually summarized as a statistical value representing the area under the ROC carve known as Area Under Curve (AUC). The AUC represents the probability that the outcome of the generated model is better than induction using a random model, where a random model has an AUC value of 0.5 while a perfect model has an AUC of 1. Therefore, the higher AUC value is, the better the model achieves.

IV. Experimental Results and Discussions

In this research study, eight machine learning algorithms were employed on the studied dataset (JIRA bug reports). These algorithms are NB, NBM, SVMs, J48, RF, LMT, JRip and KNN. As discussed before, IG feature selection was applied and features obtained after pre-processing are ranked and the top ‘N’ scoring features are selected based on ranking results. A set of eight datasets are generated, these datasets include the original dataset with the whole feature set and the top 25, 50, 75, 100, 125, 150, 200 and 300 terms are generated as new datasets to build prediction models. The results then analyzed in terms of selected performance measures, which are the accuracy, F-measure and AUC measures. These measures are used to examine the performance of machine learning algorithms empirically and the best models were reported. The open-source Weka software [45] was used to run the experiments with different parameters setting and the best parameters setting are selected based on evaluation measures.

Table II shows the results of different machine learning algorithms concerning the original JIRA dataset with all feature set generated after pre-processing. Based on the accuracy, it is found that the accuracy of different classifier varies in the range of 75.23% and 84.55%. However, the accuracy results are somehow biased as the dataset is imbalanced and the results are biased toward the larger class. The reported performance results of all algorithms based on the AUC and F-measure are promising as these measures take into consideration the performance results of all classes. As seen in Table II, the AUC performance results vary between 0.59-0.88 and F-measure between 0.85-0.90. These measures are not biased and report the performance of the two classes. Based on the F-measure performance results, all DT algorithms, namely Random Forest, LMT and J48 perform the best followed by NBM and JRip and SVM, and the KNN performed the lowest.

Table III shows the performance of different machine learning algorithms on the studied dataset with a varying number of selected features based on the IG ranging from 25-300 features. As shown in Table III, The performance results based on all measures show comparable and better results than the ones obtained from the generated models on the original dataset (with no feature selection). Therefore, the selected terms based on feature selection will able to distinguish the bug severity based on a smaller feature set. The accuracy reaches 86.54% when the number of selected features is 75 features, as compared to the model generated based on the original dataset (with all features) which was 84.55%. In terms of the AUC measure, the AUC result reaches 0.91 when the number of selected features is 125 features, as compared to the model generated based on all features which was 0.88. Furthermore, the F-measure results reached 0.91 when the number of selected features is 125 features, as compared to the model generated based on all features which was 0.9.

Fig. 3 shows the performance results of all machine learning algorithms, in terms of accuracy, on varying generated datasets based on the utilized feature selection method. According to the results, it is found that the accuracy of most algorithms stabilized when the number of selected features is 75 or more terms except NB classifier. The accuracy of NB algorithm has consistent accuracy in all datasets with a varying number of features. The performance results of AUC and F-Measure of various algorithms are shown in Fig. 4 and Fig. 5 From these figures, the performance of all classifier depends on the number of terms considered to build the classifier and the performance results provide comparable and better results than relying on the whole feature set. The reported results of the AUC and F-measure using different machine learning algorithms state that the best performance results are when the number of selected features is 125 terms. These terms are chosen using the information gain measure. It is clear that the severity of bug reports can be predicted with a reasonable performance of AUC measure as they show a differentiated and progressive value of AUC and better than using whole feature sets. Similar trends are also observed with the F-measure shown in Fig. 5 From results reported in Table III and Fig. 5, the best F-measure results were obtained using Naïve Bayes Multinomial, SVM, and Logistic model trees. These algorithms receive 90%-91% performance results where Logistic model trees achieve best and stable results. However, other algorithms report varied and fewer performances.

Similar results observed from Table III when the number of terms taken into account is over 75 terms. The AUC and F-measure results are similar and consistent as compared to the performance results when the number of selected terms are less. The results suggest that the generated models can predict the severity of bugs more accurately when the number of selected terms in the range of 75 to 125 terms. As Table III shows, it can be concluded that models have performed well in predicting the bug reports of either severity levels as reported by both AUC and f-measures values.

TABLE. II. PERFORMANCE RESULTS OF DIFFERENT CLASSIFICATION ALGORITHMS ON JIRA DATASET WITH ALL FEATURES

Classifier	**Accuracy**	**AUC**	**F-measure**
SVM	77.5	0.59	0.87
KNN	75.23	0.75	0.85
DT - J48	80.48	0.76	0.87
DT - RF	84.55	0.88	0.9
Rule-Based JRip	81.28	0.72	0.88
NBM	82.43	0.86	0.88
Naïve Bayes	80.03	0.84	0.86
DT:LMT	83.26	0.86	0.89

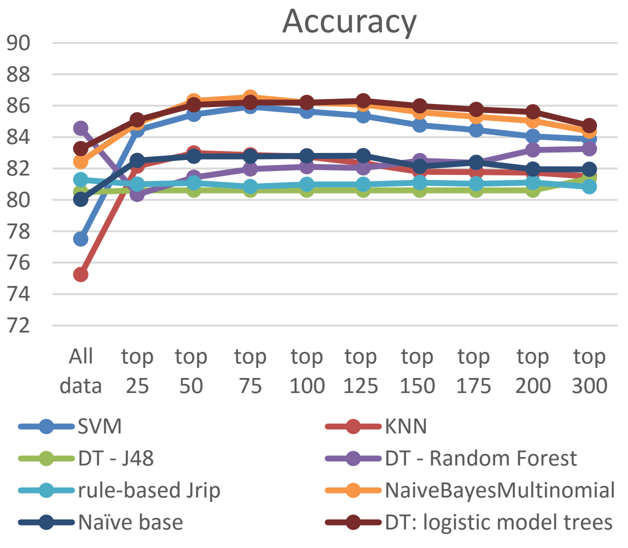


Fig. 3. Accuracy of different Classification Algorithms.

TABLE. III. ACCURACY OF DIFFERENT CLASSIFICATION ALGORITHMS WITH A VARYING NUMBER OF SELECTED FEATURES BASED ON IG

	Top 25			*Top 50*			*Top 75*		
	Accuracy	*AUC*	*F-measure*	*Accuracy*	*AUC*	*F-measure*	*Accuracy*	*AUC*	*F-measure*
SVM	84.44	0.75	0.90	85.44	0.77	0.91	85.94	0.78	0.91
KNN	82.15	0.83	0.88	82.98	0.84	0.89	82.86	0.85	0.89
DT - J48	80.61	0.74	0.88	80.61	0.74	0.88	80.61	0.74	0.88
DT - RF	80.34	0.84	0.87	81.43	0.85	0.87	81.96	0.86	0.88
rule-based JRip	80.99	0.72	0.88	81.08	0.72	0.88	80.84	0.72	0.88
NBM	84.89	0.86	0.90	86.31	0.90	0.91	86.54	0.91	0.91
Naïve Bayes	82.50	0.86	0.88	82.77	0.87	0.88	82.77	0.87	0.88
DT:LMT	85.09	0.87	0.90	86.05	0.89	0.91	86.20	0.90	0.91

	Top 100			*Top 125*			*Top 150*		
	Accuracy	*AUC*	*F-measure*	*Accuracy*	*AUC*	*F-measure*	*Accuracy*	*AUC*	*F-measure*
SVM	85.63	0.77	0.91	85.35	0.76	0.91	84.76	0.75	0.90
KNN	82.76	0.84	0.89	82.33	0.85	0.88	81.79	0.84	0.88
DT - J48	80.61	0.74	0.88	80.61	0.74	0.88	80.61	0.74	0.88
DT - RF	82.11	0.86	0.88	82.03	0.87	0.88	82.49	0.87	0.88
rule-based JRip	80.98	0.72	0.88	80.98	0.72	0.88	81.10	0.72	0.88
NBM	86.20	0.91	0.91	86.08	0.90	0.91	85.58	0.90	0.91
Naïve Bayes	82.80	0.87	0.88	82.81	0.87	0.88	82.13	0.86	0.88
DT:LMT	86.19	0.90	0.91	86.31	0.90	0.91	85.98	0.89	0.91

	Top 175			*Top 200*			*Top 300*		
	Accuracy	*AUC*	*F-measure*	*Accuracy*	*AUC*	*F-measure*	*Accuracy*	*AUC*	*F- measure*
SVM	84.45	0.74	0.90	84.04	0.73	0.90	83.87	0.72	0.90
KNN	81.77	0.84	0.88	81.74	0.83	0.88	81.49	0.81	0.88
DT - J48	80.61	0.74	0.88	80.61	0.74	0.88	81.36	0.78	0.88
DT - RF	82.36	0.87	0.88	83.18	0.87	0.89	83.25	0.87	0.89
rule-based JRip	81.02	0.72	0.88	81.11	0.72	0.88	80.83	0.72	0.88
NBM	85.30	0.90	0.90	85.03	0.90	0.90	84.37	0.88	0.90
Naïve Bayes	82.38	0.86	0.88	81.95	0.86	0.88	81.94	0.86	0.88
DT: LMT	85.75	0.89	0.91	85.60	0.89	0.91	84.73	0.88	0.90

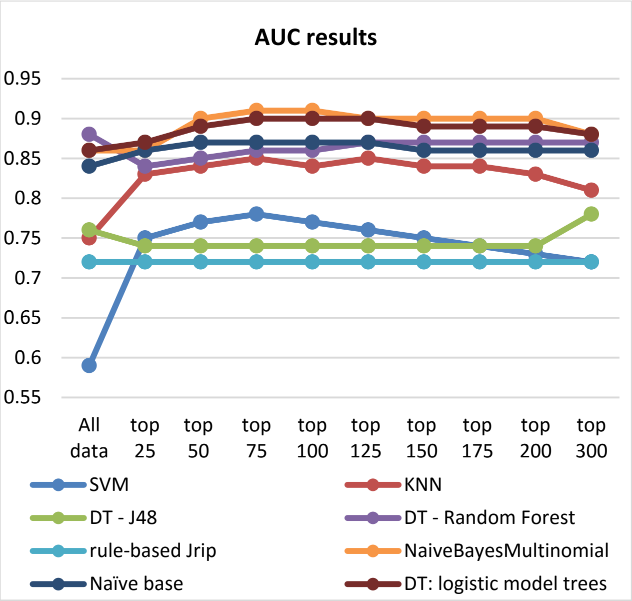


Fig. 4. The AUC Results of different Classification Algorithms with a Varying Number of Selected Terms.

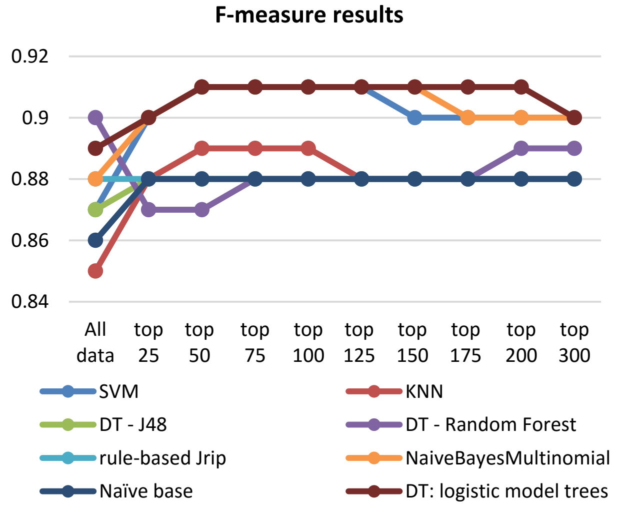


Fig. 5. The F-Measure Results of different Classification Algorithms with a Varying Number of Selected Terms.

Overall, the LMT algorithm reported the best performance results based on all performance measures. The maximum AUC and F-measure obtained with LMT are 0.90 and 0.91, respectively. It is further found that NBM performs similarly to LMT in terms of AUC and F-measure. The minimum AUC and F-measure values obtained with NBM are 0.88 and 0.91 respectively, and the maximums are 0.9 and 0.91, respectively. The SVM model also has comparable performance results, but it is in some ways less than the results reported for the LMT and NBM. The maximum AUC and F-measure results obtained with SVM are 0.8 and 0.9, respectively. Other algorithms report divergent accuracy performance results. The maximum AUC and F-measure results obtained with other algorithms are 0.86 and 0.88, respectively. Therefore, the reported results indicate that LMT performs better for bug severity prediction under the reported experimental setup. The superior performance of the LMT algorithm can be attributed to the fact that the number of selected features after applying feature selection method is relatively small compared to the original feature set. Therefore, discarding unimportant features leads LMT algorithm to build classification trees that are significantly smaller than the standard classification trees and has accurate prediction results.

Finally, the severity of new bug reports submitted to JIRA bug tracking system can be predicted automatically based on the reported prediction model. This automatic prediction would be beneficial for software engineers and developers of INTIX Company to automatically assign the severity of reported bugs instead of manual work. As a consequence, severe bugs will be solved on time without causing a delay for fixing them.

V. Conclusions

In this research paper, we described machine learning approaches for predicting the severity level of software bug reports in closed-source projects and leveraged Information Gain (IG) feature selection method to enhance the performance of the prediction by increasing the prediction accuracy of the bug severity level.

The procedure of the proposed methodology was illustrated and evaluated by comparing eight popular machine learning algorithms, namely Naive Bayes, Naive Bayes Multinomial, Support Vector Machine (SVM), Decision Tree (J48), Random Forest, Logistic Model Trees (LMT), Decision Rules (JRip) and KNN. The performances of utilized machine learning algorithms were evaluated in terms of accuracy, F-measure and Area Under the Curve (AUC). Furthermore, one of the objectives of this research work was to utilize feature selection techniques for the enhancement of selected features that will, in turn, allow for more precise discriminative power and produce better prediction performance results.

The proposed methodology was conducted on the existing severity levels assigned manually by the developers of INTIX Company through the JIRA bug tracking system. The experiments were evaluated using performance measures: accuracy, F-measure and AUC. The obtained results indicated that the Logistic Model Trees (DT: LMT) outperformed other classifiers and the overall performance has been enhanced after applying feature selection method. The results showed that the LMT algorithms reported the best performance results based on all performance measures (Accuracy= 86.31, AUC= 0.90, F-measure= 0.91).

In future work, we plan to adopt sentiment analysis and opinion mining techniques to enhance the performance of the process of predicting the severity level of software bug reports.

Acknowledgment

The authors are thankful to INTIX Company for providing us with the bug report dataset.

References

[1] T. Zhang, H. Jiang, X. Luo, and A. T. Chan, "A literature review of research in bug resolution: Tasks, challenges and future directions," The Computer Journal, vol. 59, pp. 741-773, 2016.

[2] C. Liu, J. Yang, L. Tan, and M. Hafiz, "R2Fix: Automatically generating bug fixes from bug reports," in 2013 IEEE Sixth International Conference on Software Testing, Verification and Validation, 2013, pp. 282-291.

[3] G. Murphy and D. Cubranic, "Automatic bug triage using text categorization," in Proceedings of the Sixteenth International Conference on Software Engineering & Knowledge Engineering, 2004.

[4] N. Bettenburg, S. Just, A. Schröter, C. Weiss, R. Premraj, and T. Zimmermann, "What makes a good bug report?," in Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering, 2008, pp. 308-318.

[5] Bugzilla. (2019). Bug Tracking System. Available: http://www.bugzilla.org/

[6] Atlassian. (2019). JIRA. Available: http://www.atlassian.com/ software/jira/

[7] G. Antoniol, K. Ayari, M. Di Penta, F. Khomh, and Y.-G. Guéhéneuc, "Is it a bug or an enhancement?: a text-based approach to classify change requests," in CASCON, 2008, pp. 304-318.

[8] J. Uddin, R. Ghazali, M. M. Deris, R. Naseem, and H. Shah, "A survey on bug prioritization," Artificial Intelligence Review, vol. 47, pp. 145-180, 2017.

[9] K. Chaturvedi and V. Singh, "Determining bug severity using machine learning techniques," in 2012 CSI Sixth International Conference on Software Engineering (CONSEG), 2012, pp. 1-6.

[10] A. Lamkanfi, S. Demeyer, Q. D. Soetens, and T. Verdonck, "Comparing mining algorithms for predicting the severity of a reported bug," in 2011 15th European Conference on Software Maintenance and Reengineering, 2011, pp. 249-258.

[11] A. Lamkanfi, S. Demeyer, E. Giger, and B. Goethals, "Predicting the severity of a reported bug," in 2010 7th IEEE Working Conference on Mining Software Repositories (MSR 2010), 2010, pp. 1-10.

[12] I. Herraiz, D. M. German, J. M. Gonzalez-Barahona, and G. Robles, "Towards a simplification of the bug report form in eclipse," in Proceedings of the 2008 international working conference on Mining software repositories, 2008, pp. 145-148.

[13] Y. Tian, D. Lo, and C. Sun, "Information retrieval based nearest neighbor classification for fine-grained bug severity prediction," in 2012 19th Working Conference on Reverse Engineering, 2012, pp. 215-224.

[14] T. Zhang, J. Chen, G. Yang, B. Lee, and X. Luo, "Towards more accurate severity prediction and fixer recommendation of software bugs," Journal of Systems and Software, vol. 117, pp. 166-184, 2016.

[15] T. Menzies and A. Marcus, "Automated severity assessment of software defect reports," in 2008 IEEE International Conference on Software Maintenance, 2008, pp. 346-355.

[16] R. Jindal, R. Malhotra, and A. Jain, "Prediction of defect severity by mining software project reports," International Journal of System Assurance Engineering and Management, vol. 8, pp. 334-351, 2017.

[17] A. F. Otoom, D. Al-Shdaifat, M. Hammad, and E. E. Abdallah, "Severity prediction of software bugs," in 2016 7th International Conference on Information and Communication Systems (ICICS), 2016, pp. 92-95.

[18] K. Jin, A. Dashbalbar, G. Yang, B. Lee, and J.-W. Lee, "Improving predictions about bug severity by utilizing bugs classified as normal," Contemp Eng Sci, vol. 9, pp. 933-942, 2016.

[19] C.-Z. Yang, K.-Y. Chen, W.-C. Kao, and C.-C. Yang, "Improving severity prediction on software bug reports using quality indicators," in 2014 IEEE 5th International Conference on Software Engineering and Service Science, 2014, pp. 216-219.

[20] V. Singh, S. Misra, and M. Sharma, "Bug severity assessment in cross project context and identifying training candidates," Journal of Information & Knowledge Management, vol. 16, p. 1750005, 2017.

[21] W. Liu, S. Wang, X. Chen, and H. Jiang, "Predicting the severity of bug reports based on feature selection," International Journal of Software Engineering and Knowledge Engineering, vol. 28, pp. 537-558, 2018.

[22] N. K. S. Roy and B. Rossi, "Towards an improvement of bug severity classification," in 2014 40th EUROMICRO Conference on Software Engineering and Advanced Applications, 2014, pp. 269-276.

[23] G. Sharma, S. Sharma, and S. Gujral, "A novel way of assessing software bug severity using dictionary of critical terms," Procedia Computer Science, vol. 70, pp. 632-639, 2015.

[24] S. Sharmin, F. Aktar, A. A. Ali, M. A. H. Khan, and M. Shoyaib, "Bfsp: A feature selection method for bug severity classification," in 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017, pp. 750-754.

[25] C.-Z. Yang, C.-C. Hou, W.-C. Kao, and X. Chen, "An empirical study on improving severity prediction of defect reports using feature selection," in 2012 19th Asia-Pacific Software Engineering Conference, 2012, pp. 240-249.

[26] G. Yang, K. Min, J.-W. Lee, and B. Lee, "Applying Topic Modeling and Similarity for Predicting Bug Severity in Cross Projects," KSII Transactions on Internet & Information Systems, vol. 13, 2019.

[27] G. Yang, T. Zhang, and B. Lee, "An emotion similarity based severity prediction of software bugs: A case study of open source projects," IEICE TRANSACTIONS on Information and Systems, vol. 101, pp. 2015-2026, 2018.

[28] W. Y. Ramay, Q. Umer, X. C. Yin, C. Zhu, and I. Illahi, "Deep Neural Network-Based Severity Prediction of Bug Reports," IEEE Access, vol. 7, pp. 46846-46857, 2019.

[29] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation on feature selection for text clustering," in Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 488-495.

[30] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in Icml, 1997, p. 35.

[31] M. F. Caropreso, S. Matwin, and F. Sebastiani, "A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization," Text databases and document management: Theory and practice, vol. 5478, pp. 78-102, 2001.

[32] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining: predictive methods for analyzing unstructured information: Springer Science & Business Media, 2010.

[33] P. S. Laplace, "Memoir on the probability of the causes of events," Statistical Science, vol. 1, pp. 364-378, 1986.

[34] T. M. Mitchell, "Machine Learning," ed: Mcgraw-hill, 1997.

[35] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in AAAI-98 workshop on learning for text categorization, 1998, pp. 41-48.

[36] V. Christlein, D. Bernecker, F. Hönig, A. Maier, and E. Angelopoulou, "Writer identification using GMM supervectors and exemplar-SVMs," Pattern Recognition, vol. 63, pp. 258-267, 2017.

[37] S. Cogill and L. Wang, "Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates," Bioinformatics, vol. 32, pp. 3611-3618, 2016.

[38] F. Simistira, V. Katsouros, and G. Carayannis, "Recognition of online handwritten mathematical formulas using probabilistic SVMs and stochastic context free grammars," Pattern Recognition Letters, vol. 53, pp. 85-92, 2015.

[39] D. N. Sotiropoulos, D. E. Pournarakis, and G. M. Giaglis, "SVM-based sentiment classification: a comparative study against state-of-the-art classifiers," International Journal of Computational Intelligence Studies, vol. 6, pp. 52-67, 2017.

[40] J. Platt, "Fast training of support vector machines using sequential minimal optimization, in, B. Scholkopf, C. Burges, A. Smola,(eds.): Advances in Kernel Methods-Support Vector Learning," ed: MIT Press, 1998.

[41] J. R. Quinlan, C4. 5: programs for machine learning: Elsevier, 2014.

[42] L. Breiman, "Random forests," Machine learning, vol.45,pp. 5-32, 2001.

[43] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," Machine learning, vol. 59, pp. 161-205, 2005.

[44] W. W. Cohen, "Fast effective rule induction," in Machine Learning Proceedings 1995, ed: Elsevier, 1995, pp. 115-123.

[45] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann, 2016.

Steganography Performance over AWGN Channel

Fahd Alharbi
Faculty of Engineering
King Abdulaziz University, Rabigh, Saudi Arabia

***Abstract*—Steganography can be performed using frequency domain or spatial domain. In spatial domain method, the least significant bits (LSB) is the mostly used method where the least significant bits of the image's pixels binary representation are used to carry the confidential data bits. On the other hand, secret data bits in the frequency domain technique are hidden using coefficients of the image frequency representation such as discrete cosine transform (DCT). Robustness against image attacks or channel's noise is a key requirement in steganography. In this paper, we study the performance of the steganography methods over a channel with Added White Gaussian Noise (AWGN). We use the bit error rate to evaluate the performance of each method over a channel with different noise levels. Simulation results show that the frequency domain technique is more robust and achieves better bit error rate in a noisy channel than the spatial domain method. Moreover, we enhanced the steganography system robustness by using convolution encoder and Viterbi decoder. The effect of the encoder's parameters, such as rate and constraint length is evaluated.**

Keywords*—*Steganography; robustness; noise; AWGN; viterbi

I. Introduction

Steganography is the process of concealing critical and important information undetectably in a cover medium such as image, voice and video [1, 2]. The steganography model is illustrated in Fig. 1, where the important message hidden in the cover image using the embedding process. The watermarked image which is the cover image with concealed data is transmitted through a communication channel. The receiver recovers the secret message using the extracting process. The embedding and extracting processes are implemented using spatial domain [3-5] or frequency domain techniques [6, 7].

A. *Spatial Domain Steganography*

The mostly used method for spatial domain steganography is the least significant bit (LSB). LSB first decomposes the image's pixels value using the binary system, then insert the secret bits into the least significant bits. For example, the process of hiding the letter A (ASCII Code: 1000001) into seven pixels of a gray-scale image is illustrated at Table I. The effect of the LSB on the cover image is hardly noticed by the human eye since the bit used for data hiding has a small value.

B. *LSB Performance*

Now, we study the performance of the LSB embedding method in a communication system with Added White Gaussian Noise channel (Fig. 2). The watermark image W is hidden into the cover image C using the LSB embedding process module. The watermarked image I is transmitted through an AWGN communication channel. The received watermarked image I' is fed to the LSB extracting process module to extract the recovered watermark image W'.

We change the variance of Added White Gaussian Noise for different channel noise levels. The effect on the watermarked image I is evaluated by computing PSNR (Peak Signal to Noise Ratio). The PSNR between the 8-bit gray-scale watermarked image I and the received watermarked image I' is computed as following:

$$PSNR = \frac{10\log_{10}(255)^2}{\frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left[I(i,j)-I'(i,j)\right]^2} \quad (1)$$

where M and N represent the size the watermarked image.

The performance of the LSB steganography over the AWGN channel is measured by computing the bit error rate (BER). The BER is calculated using Eq. (2), where the watermark image W is a black-and-white image with a size of K by L pixels. The calculation performs Exclusive-OR ($\oplus$) between the watermark bits $W(i,j)$ and the recovered watermark bits $W'(i,j)$.

$$BER = \frac{\sum_{i=0}^{K-1}\sum_{j=0}^{L-1}\left[W(i,j)\oplus W'(i,j)\right]}{KL} \quad (2)$$

The watermarked image I (Fig. 3) is generated by hiding the watermark image W into the cover image C using the LSB embedding technique. The watermarked image is transmitted through the AWGN channel. We vary the AWGN level and for each case, we compute the PSNR and the BER. The performance of the LSB steganography is illustrated at Table II and Fig. 4, where LSB is very weak against the channel noise.

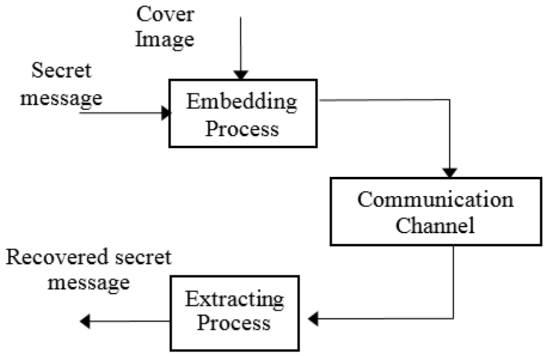


Fig. 1. Steganography Model.

TABLE. I. LSB EMBEDDING

Pixel number	Pixels before embedding	letter A	Pixels after embedding	Effects on pixel value
1st	01001111	1	0100111**1**	0
2nd	01001101	0	0100110**0**	-1
3rd	01001110	0	0100111**0**	0
4th	01001111	0	0100111**0**	-1
5th	01010000	0	0101000**0**	0
6th	01001011	0	0100101**0**	-1
7th	01010000	1	0101000**1**	+1

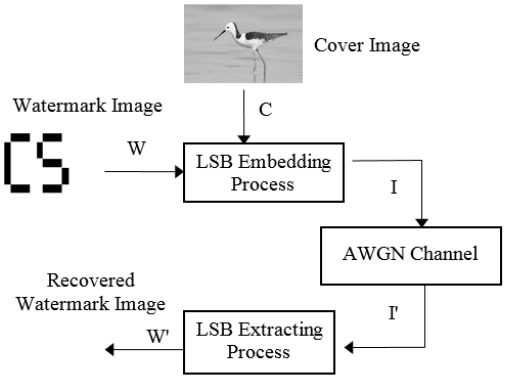
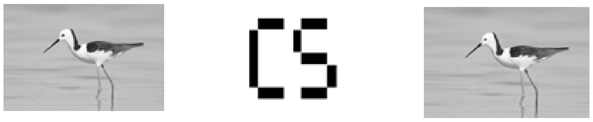


Fig. 2. LSB Steganography Model.



(a) Cover Image C (b) Watermark Image W (c) Watermarked Image I

Fig. 3. Embedding Process.

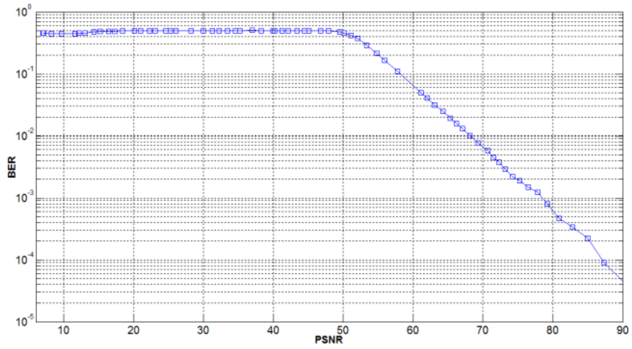


Fig. 4. LSB Steganography Performance.

C. Frequency Domain Steganography

In the frequency domain methods, the cover image is first transformed from spatial to frequency domain using the discrete cosine transform (DCT) and then the secret message bits are embedded using the transformation coefficients. The cover image C is divided into 8 x 8 blocks *f(x,y)* where each block is converted using the 2D DCT (Eq. 3) to its Discrete Cosine Transform *F(x,y)* coefficient blocks.

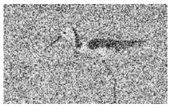
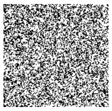
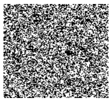
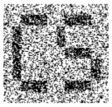
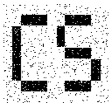
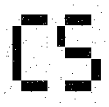
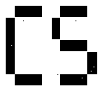
$$F(u,v) = \alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right] \tag{3}$$

where

$$\alpha(k) = \begin{cases} \sqrt{\dfrac{1}{N}} & \text{for k} = 0 \\ \sqrt{\dfrac{2}{N}} & \text{for k} = 1, 2, \ldots. N-1 \end{cases}$$

The 2D DCT transformation result in 8 x 8 matrix of DCT2 coefficients is shown at Fig. 5, where the frequency components are the lowest frequency components FL, the middle frequency components FM and the higher frequency components FH [8]. Embedding in the FL components will affect the cover image quality while embedding in the FH components will make the secret message vulnerable against the lossy compression. Accordingly, the FM components are selected to embed the watermark to maintain the cover image quality and provide resistance to the compression. The secret message bit then embedded into the 2D DCT coefficients and the inverse 2D DCT is performed. This process is repeated for the other secret bits where each bit embedded in an 8 x 8 block.

TABLE. II. LSB STEGANOGRAPHY PERFORMANCE

PSNR	Received Watermarked Image I'	Recovered watermark W'	BER
10 dB			0.45
50 dB			0.45
53 dB			0.3
60 dB			0.06
70 dB			0.006
80 dB			00006

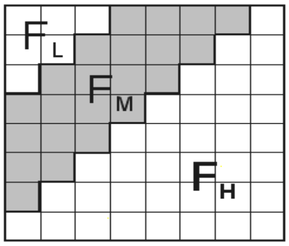


Fig. 5. The 2D DCT Frequency Components.

D. DCT Performance

The process of the DCT embedding method in a communication system with Added White Gaussian Noise channel is shown at Fig. 6. The watermark image W is hidden into the cover image C using the DCT embedding process module. The watermarked image I is transmitted through an AWGN communication channel. The received watermarked image I' is fed to the DCT extracting process module to extract the recovered watermark image W'.

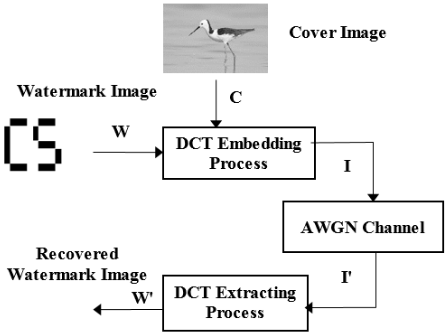


Fig. 6. DCT Steganography Model.

The watermarked image I (Fig. 7) is generated by hiding the watermark image W into the cover image C using the DCT embedding technique. The performance of the DCT steganography is illustrated at Table III and Fig. 8, where the DCT steganography technique is more robust against the channel noise than the LSB steganography technique.

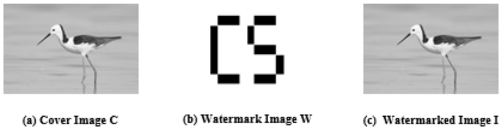


Fig. 7. DCT Embedding Process.

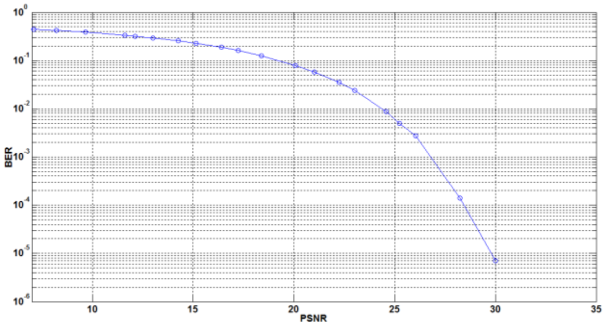
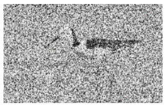
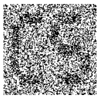
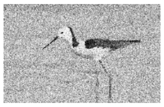
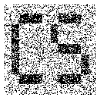
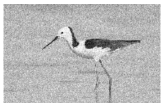
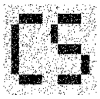
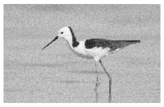
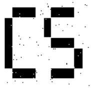


Fig. 8. DCT Steganography Performance.

TABLE. III. DCT STEGANOGRAPHY PERFORMANCE

PSNR	Received Watermarked Image I'	Recovered watermark W'	BER
10 dB			0.37
15 dB			0.23
20 dB			0.08
25 dB			0.0053
28 dB			0.00013
30 dB			0.000007

II. ROBUST STEGANOGRAPHY SYSTEM

In this section, the robustness of the steganography system is enhanced by introducing the errors detection and correction capability (Fig. 9) [9-12]. The watermark bits are encoded using convolutional encoder (Fig. 10), where each m-bit at the encoder input are encoded into n-bit symbol. The convolutional encoder is categorized by the code rate $R = m / n$ and the constraint length CL that represent the number of memory elements. The encoded watermark bits are fed to the embedding process at the transmitter. At the receiver side, the received bits ate coded using a Viterbi decoder to obtain the watermark W'.

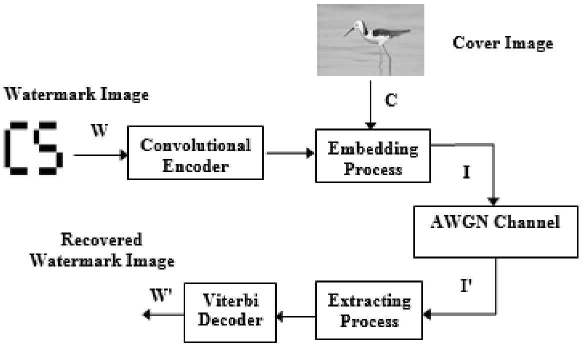


Fig. 9. Robust Steganography Model.

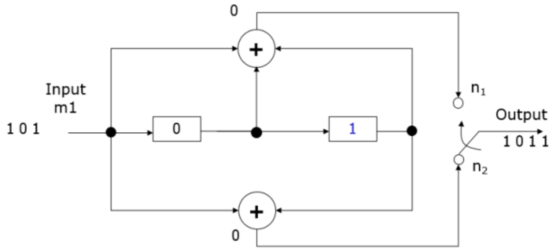


Fig. 10. Convolutional Encoder *R* =1/2 and *CL* =3.

A. *Robust Steganography Performance*

The performance of the robust steganography system is evaluated for the LSB embedding method and the DCT embedding method. The experiments evaluate the impact of convolutional encoder parameters such as the code rate *R* and the constraint length *CL*. The performance of the LSB steganography using convolutional encoder and Viterbi decoder is illustrated at Fig. 11-14. The encoder rate *R* is set to 1/2 and we vary the constraint length *CL* (Fig. 11). Moreover, the constraint length *CL* is set and we vary the encoder rate *R* (Fig. 12-14) where it is clear that the encoded LSB technique is still weak against the AWGN channel. On the other hand, the performance of the DCT steganography using convolutional encoder and Viterbi decoder is illustrated at Fig. 15-18 where the robustness is significantly enhanced. The results show that with higher *R* and *CL*, the system will achieve better robustness against the AWGN channel.

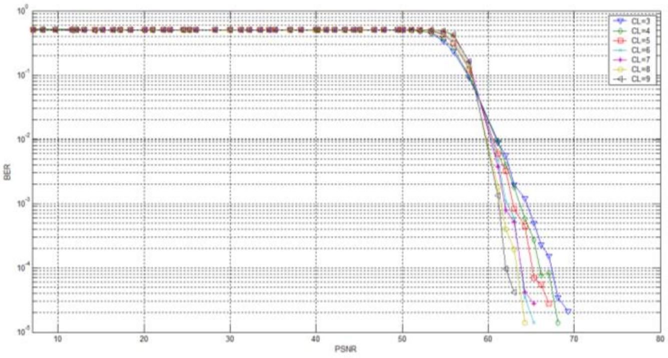


Fig. 11. LSB Steganography Performance with R=1/2.

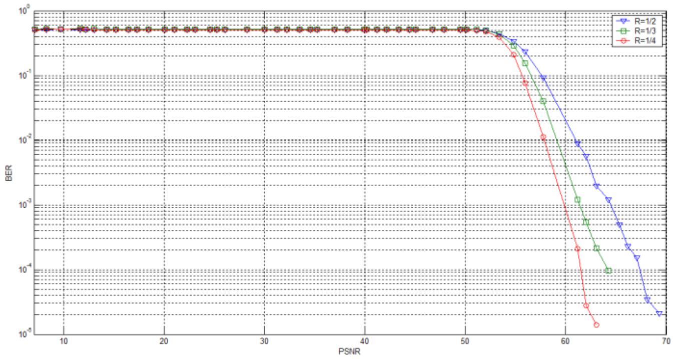


Fig. 12. LSB Steganography Performance with CL=3.

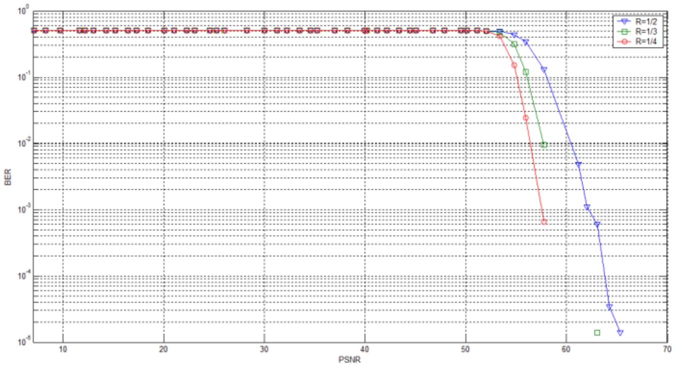


Fig. 13. LSB Steganography Performance with CL=6.

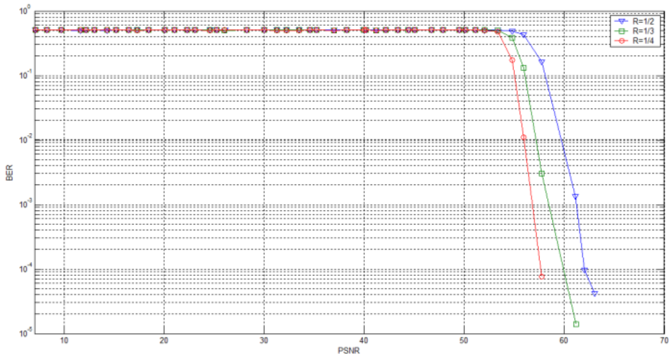


Fig. 14. LSB Steganography Performance with CL=9.

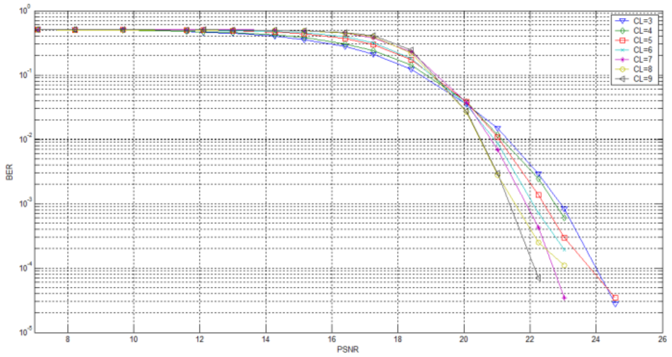


Fig. 15. DCT Steganography Performance with R=1/2.

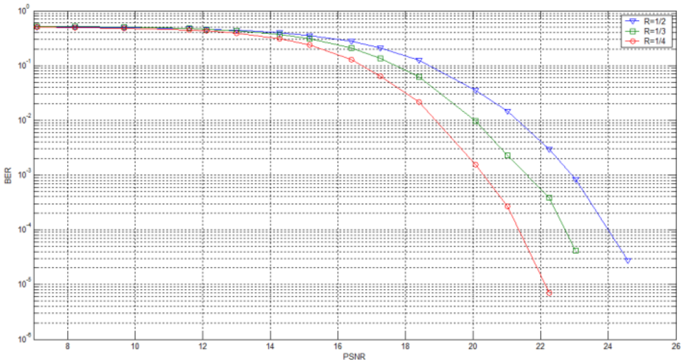


Fig. 16. DCT Steganography Performance with CL=3.

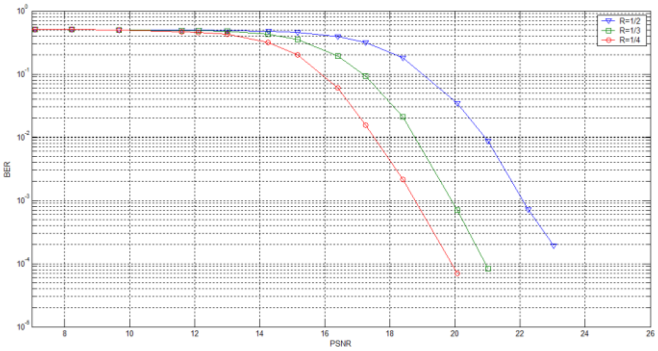


Fig. 17. DCT Steganography Performance with CL=6.

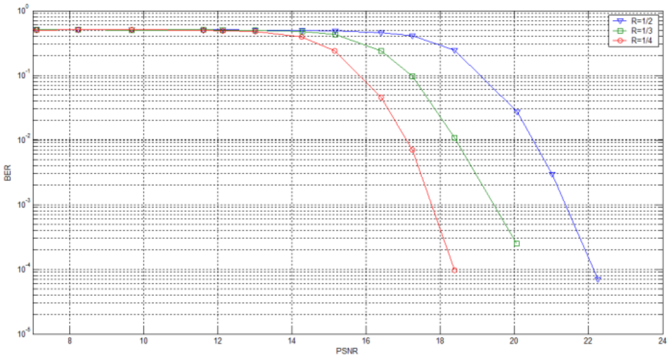


Fig. 18. DCT Steganography Performance with CL=9.

III. Conclusion

In this paper, steganography performance over a channel with Added White Gaussian Noise a noisy is evaluated. The secret message concealed using frequency domain or spatial domain then transmitted over the noisy channel. The performance of the steganography technique is measured by computing the bit error rate for each method over a channel with different noise levels. The results show that the DCT technique outperform the LSB technique in achieving better BER. Also, the steganography robustness is enhanced by using convolution encoder and Viterbi decoder. Moreover, the impact of the encoder's parameters such as rate and constraint length are evaluated. Simulation results illustrate that the steganography system robustness against the AWGN channel is improved with higher *R* and *CL*.

References

[1] Fridrich, J., [Steganography in Digital Media: Principles, Algorithms, and Applications], Cambridge University Press (2009).

[2] Mansi S. Subhedar a, Vijay H. Mankar b," Current status and key issues in image steganography: A survey," Computer Science Review Volumes 13–14, November 2014, Pages 95-113.

[3] M. Gaaed and M. Tahar, "Digital Image Watermarking based on LSB Techniques: A Comparative Study", International Journal of Computer Applications, vol. 181, no. 26, pp. 30-36, 2018.

[4] H. Zangana, "Watermarking System Using LSB", IOSR Journal of Computer Engineering, vol. 19, no. 3, pp. 75-79, 2017.

[5] Fahd Alharbi," Novel Steganography System using Lucas Sequence," International Journal of Advanced Computer Science and Applications(IJACSA), Volume 4 Issue 4, 2013.

[6] H. Fang and Z. Hua, "A Study on the Performance of Watermarking Algorithm Based on DCT", Advanced Materials Research, vol. 846-847, pp. 1040-1043, 2013.

[7] D. Singh and S. Singh, "DCT based efficient fragile watermarking scheme for image authentication and restoration", Multimedia Tools and Applications, vol. 76, no. 1, pp. 953-977, 2015.

[8] Amin P.K., Ning Liu, Subbalakshmi K. P. "Statistical Secure Digital Image Data Hiding", IEEE 7th workshop on Multimedia Signal Processing .pp.1-4, 2005.

[9] S. V. Viraktamath1 , Preeya H. Patil, G. V. Attimarad "Impact of code rate on the performance of Viterbi decoder in AWGN channel", 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014.

[10] P. Elias, Predictive coding--II, IRE Transactions on Information Theory, 1955. vol. 1, no. 1, p. 24 - 33.

[11] SKLAR, B. Digital Communications: Fundamentals and Applications. 2nd ed. System View, 2001.

[12] BATSON, B. H., MOOREHEAD, R. W. Simulation Results for the Viterbi Decoding Algorithm. NASA-TR-R-396, Technical report, 1972.

Cloud Security based on the Homomorphic Encryption

Waleed T. Al-Sit[1], Hani Al-Zoubi[3]
Department of Computer Engineering
Mu'tah University, Al-Karak, Jordan

Qussay Al-Jubouri[2]
Department of of Communication Engineering
University of Technology, Baghdad, Iraq

***Abstract*—Cloud computing provides services rather than products; where it offers many benefits to clients who pay to use hardware and software resources. There are many advantages of using cloud computing such as low cost, easy to maintain, and available resources. The main challenge in the Cloud system is how to obtain a highly secured system against attackers. For this reason, methods were developed to increase the security level in different techniques. This paper aims to review these techniques with their security challenges by presenting the most popular cloud techniques and applications. Homomorphic Encryption method in cloud computing is presented in this paper as a solution to increase the security of the data. By using this method, a client can perform an operation on encrypted data without being decrypted which is the same result as the computation applied to decrypted data. Finally, the reviewed security techniques are discussed with some recommendations that might be used to raise the required security level in such a system.**

Keywords—Cloud computing; homomorphic encryption; security

I. Introduction

Cloud computing represents the latest effort in delivering computing resources as a service. It enables any organization to obtain its computing resources and applications from any location via an internet connection. As reported in [1], the US National Institute of Standards and Technology (NIST) have defined cloud computing as "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources [2]. Networks, servers, storage, applications, and services can be rapidly provisioned and released with minimal management effort or service provider interaction".

In Cloud, computing Clients are permitted to store a vast amount of information on distributed storage, and it provides on-demand services over a network, by payment method. Different security issues like confidentiality, privacy, authentication, and integrity needed to address. The more significant part of the cloud administration supplier stores the information in plaintext and client need to utilize their encryption algorithm to secure their information. The information should be decrypted when t it is to be computed. As data is decrypted, it will be prone to attack. Therefore, finding a solution is essential to protect the data, and to perform any computation on it without decryption; this process is called Homomorphic Encryption. Several advantages of using cloud computing such as (i) low cost, (ii) easy to maintain, (iii) backup personalization and recovery, and (iv) remote access [5]. However, higher operational price, security, and privacy represent the main disadvantages. In this work, Homomorphic Encryption method is presented as a solution to increase the security of the data. By using this method, a client can perform an operation on encrypted data without being decrypted which is the same result as the computation applied to decrypted data. Organization of the remaining sections of this paper is as follows. Section II presents the definition of cloud computing. Next, security issues of cloud computing are described in Section III. The challenges that face security are then presented in Section IV. After that, several applications and vulnerabilities of the cloud computing are described in Section V. Recent related works are the presented in Section VI. Next, Homomorphic Encryption is described in Section VII. Finally; the work is concluded in Section VIII.

II. Cloud Computing Definition

Cloud computing system is split into two parts; the front end and the back end [3]. Front end represents the user, while the back edge represents the service provider and the network between them can be performed as shown in Fig. 1.

Cloud computing provides the following services:

- Platform as a Service PaaS, in which cloud computing provides a platform and also an application can be built, test, and deploy. This permits the clients to manage, run, and develop new applications or services. There is no need to buy any software but need to pay for the time that uses only [4].
- A service SaaS, Cloud-Computing Software supplies licensed application to clients, such as Google Apps, and web-based emails, CRM and ERP systems.
- Infrastructure as a service IaaS, Cloud Computing on this method, provides an entire virtual data center of resource such as servers, server space, processors.

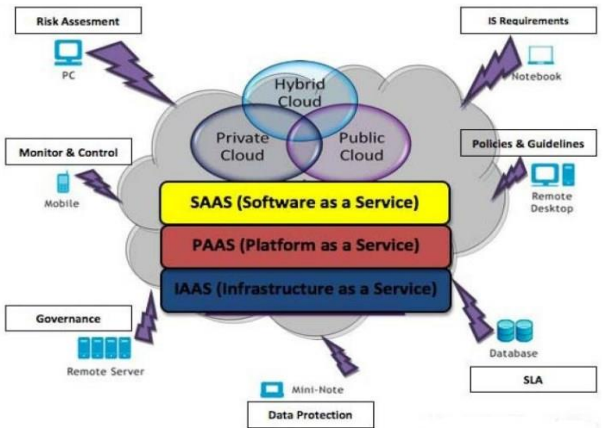


Fig. 1. Cloud Architecture [16].

IaaS hosts PaaS and SaaS, in this service, any breaking on IaaS will affect the security of both SaaS and PaaS services. If any cloud service is attacked, that will affect the other layer. The relationship between cloud models could be a source of risks. A SaaS provider could rent a development area from a PaaS provider, which may also rent infrastructure from an IaaS provider. Every provider is accountable for securing its services, which could result in an inconsistent combination of security models. It also makes confusion over which service provider is responsible if an attack happens. Utility Cloud computing is the using of computing resources by sharing it with several clients. There are a number of companies that support these services, such as Google, Amazon, and Microsoft, etc. Cloud computing is classified into four distributed models:

1) Private cloud, which is also named an internal Cloud, this type is used and modified only by the members of an individual organization. It represents a way to allow exclusive access for resources, facilities, applications, and services from everyone in a particular organization, and the organization or third-party provider owns the cloud infrastructure.

2) Public cloud, this type is made available to the industry or public where the cloud infrastructures, services, resources are available free to the public. Using this model, the services provider organization owns the Cloud (e.g., Amazon cloud service).

3) Community cloud, this type is shared among several organizations, and managed by them. By using this model, more than one organization is allowed to use and access cloud infrastructures, resources, facilities, and services. The main point of this model is that multiple organizations have the same policy and missions of work.

4) Hybrid cloud, this type combines between two or more of other computing cloud types where some services are allowed to public, where the other is granted only to an individual organization. The cloud infrastructure, services, applications, and data are a combination of private, public, and community cloud, on this module. The existence of a specific technology is necessary to allow portability (e.g., data stored in private manipulated by application in the community or public Cloud).

- Furthermore, three main components of cloud computing from the end-user perspective [7] are described as follows.
- *End-user*: Is a software or device that clients can use to access the services that are provided by the cloud service provider.
- *Cloud Network*: Represents a group of network devices that are used to connect the client with a cloud computing provider.
- *Cloud Application Programming Interface (APIs)*: A set of instructions that help programmer to develop different cloud services with various end-users.

The essential characteristics of cloud computing can be presented as five main keys below [3]:

- *On-demand self-service*: The services can be requested and managed from the cloud without the interaction with the service provider. The provision of the computing capabilities is accomplished when required automatically.
- *A broad network access*: The standard mechanism used to enable the user services and application to be accessible to the customers. The availability of the services should be heterogeneous using thin and thick clients.
- *Resource pooling*: The resources are shared to serve different costumers using multi-tenancy model. The mapping between the physical and the virtual resources provided to the end-user.
- *Rapid elasticity*: The resources are scaled-down and up as required. The current service matches the available resources.
- *Measured service*: Up-down scaling for the resources is automatically performed as well as controlling the resource usage by provisioning a metering ability for different kinds of services provided by the cloud.

Moreover, the data security of cloud computing mainly depends on the following requirements:

- *Data confidentiality*: It means that authorized people are only allowable to access data.
- *Data integrity*: It means that only authorized users perform modification on data.
- *Data availability*: It means that authorized users can access their data at any time and from anywhere. There are three main threats to availability; cloud service provider availability, network-based attack, and backup of saved data by the cloud service provider.
- *Data remanence*: This issue appears when data is removed it may be disclosed to an unauthorized part. Care must be considered when information is deleted. A simple schematic diagrams shown in Fig. 2 presents the life cycle of data.

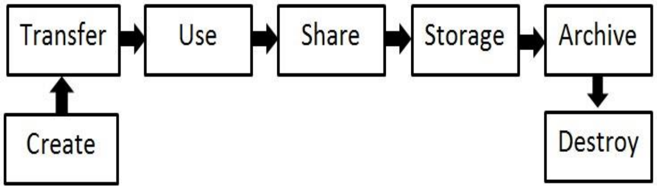


Fig. 2. Data Life Cycle [19].

III. Security Issues

Cloud-computing acceptance is affected by a low level of cloud security. It is a crucial factor to guarantee the success of any system, providing protection is a significant concern, cloud computing, which is also on this challenge.as shown in Fig. 3. However, security becomes more complicated when dealing with cloud environments and by multiple organizations that share resources.

In general, computer security focuses on these main objectives as reported in [4]. These objectives can be listed as follows:

1) Availability: It means clients can access services and resources that are provided by cloud computing providers from anywhere at any time.

2) Confidentiality: Data for the clients should be at a high level of security where it allows the authorized people to access only.

3) Data integrity: Keeps data safe, data shouldn't be lost or altered from unauthorized people or when it is stored or transport over the network.

4) Access Management (AM): Who can access the systems that have client's information? What can they access? Is the access appropriate? The answer is "provider has active identity management".

5) Audit: Everything in the system should be audited and checked, we can do that by adding a layer over the virtualized OS.

6) Control: Cloud computing provider sets strategies and regulations to organize the usage of the applications, services, resources, and so on.

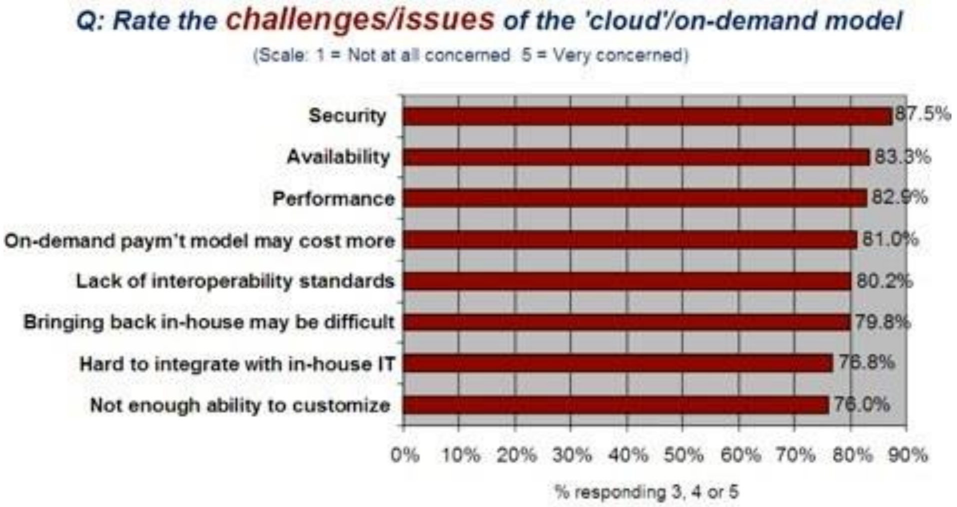


Fig. 3. Results of IDC Survey Ranking Security Challenges [8].

IV. Security Challenges

A. Delivery Model Security Challenges

The three service models are mentioned previously, and the relationship and dependency should be known between the delivery models and security challenges; to understand more about security concerns and security challenges. These layers are cumulated; which means an attack impacts on the Iaas layer, the attack will affect to the above two layers (SaaS-PaaS) [12]. The security challenge has a different level of each model and it can be summarized as follows.

In SaaS model, security challenge may raise according to the security responsibility which lies with the provider. Where, clients have less control over security when compared with two other models. So, it becomes difficult for the client to trust the security level and, also difficult to get confirmation of the services and application. In PaaS model, the clients have the ability to build their applications on the platform offered by the provider, so two security layers on this model the first one is the security of the client's application implement on the platform that is powered by the clients while the second one is the security of the platform itself, the providers are responsible for this layer. Comparing this model with SaaS model, it seems more extensible, but any security under the application platform level will be powered by the provider. Some of the PaaS challenges related to data and other issues such as; Infrastructure security and third-party –services and tools. Best security level can be achieved with the IaaS model where the clients have full control over the security such as organization configures, control infrastructure, and policy security [6]. They should control the software running in VMM, but cloud providers still manage on the underlying infrastructure.

The security level of a network is also affected by network techniques which increase the probability to deal with certain attacks as follows [9][10]:

- *Hyper Visor Attack*: On virtualization hypervisor allows more than one operating systems use the same platform hardware. This may decrease the security on each such systems since it will be difficult to track and detect the security issues (e.g., any threats occur on the guest OS may be effective to the host OS).

- *Denial of Services*: Prevent the authorized clients to access their data and applications and also because the system to slow down by a huge number of requested by attacker this causes unavailable service for the authorized people.

- *Sniffer Attacks*: A sniffer is an application that can be used to catch the packets on the network. Sniffer attacks use this application by the hacker to read data on the packet if the data is being transferred through these packets are not encrypted. So the hacker can read clients passwords, sensitive data, and so on.

- *Reused IP addresses*: Each user on the network has an IP address assume that user moves out from the network then his IP address will give to another customer. From a provider's perspective, there is no problem; since it has a limited IP address. From a user's perspective, it will create a security risk to the new user since as there is a certain time delay between the change of an IP address in DNS and the clearing of that address in DNS caches, sometimes the old IP address still has a chance to access the data. This is violating the privacy of the earlier user.

- *Google Hacking Attack*: Search Engines, such as "Google" has been used by hackers to find system security vulnerability they decide to hake. There are two types of vulnerability found on the Internet, miss configurations and software vulnerability

The security is also affected by the location of the cloud systems [10] according to three main reasons:

1) Transfer data across countries borders.
2) Various location and services provider.
3) Data collection and Mixing.

V. Cloud Applications and Vulnerabilities

A. Netflix

Netflix is a global provider online video stream and now has over 75 million subscribers, In August 2008, the concept of the Cloud applied on this application, Netflix is considered one of the largest cloud services, and it gives services as SaaS module, but it was discovered that hackers tried to use Netflix according to the weakness in silver light Something like adobe flash. These attackers use fake advertisements to install virulent content on the users' system.

B. LinkedIn

It is a social networking site designed especially for the business community. To allow members to create their profiles and interconnect with trusted people, users create their profiles, and each account will have a user name and password. In 2012, this application lost a $5 million after hacker shared 6.5 million of its hashed passwords on password cracking forum.The company takes many procedures to increase security because they emailed the members to reset their passwords, Emailverfication, CAPTCHA (public Turing test), two-step verification(Two verification is needed in order to access an account).

C. iCloud

iCloud is a cloud sServices it allows apple client systems to store their information automatically, images up to date across. It was started 2011, by Apple Inc. Hackappcom shared a method that they can guess username and passwords for the clients using app API, DEFCON group it is also (Script Information), DEFCON group believes that the reason causing the cyber –an attack is iCloud. Response to this attacks iCloud now requested from the clients to enter passwords and anther such as SMS or email alternative verification.

D. Dropbox

It is a cloud service that offers the place for the clients to store their photos, file, videos, and allows safe back up and also many facilities. It was started in 2007. In 2011, hundreds of usernames and passwords of Dropbox clients were hacked. Also, there was a programming error that allowed access to the user accounts without passwords between 1:54 PM and 5:46 PM this error was detected immediately by Dropbox. To increase security, Dropbox uses two factors verification to allow clients to access their accounts.

VI. Related Works

Many methods and algorithms used to increase security in the cloud environment. Few of which are as follows:

A. Cloud Security Tools

Recently, there are many tools that are used by Cloud to prevent an attacker accessing the client's network; here are some of the most popular security tools [11]:

- *Silver Sky*: It is a Cloud-based on email and network security solutions. It provides email auditing and monitoring.

- *DocTracker*: Offers security Control for documents on the Cloud and services are working with file sharing (e.g. Microsoft SharePoint, Box, and API for integration). Without DocTracker when you send a document out of your system you cannot follow and track it, but DocTracker allows you to follow, control, and give each user you share a document with him a specific privilege.

- *Proof point*: Offers a high speed of response to block and detect spam's spread by email and insuring relevant data that come in and come out are secured.

- *Qualys*: Offers Web Application Security, Vulnerability Management, App Security Management, Web Access Management.

- *White Hat*: Offers a high level of protection website during the coding process.

- *Valuation:* Any data transferred from your network should be encrypted; this tool provides data encryption by AES Method.

B. *Using Mobile One-Time Password (OTP)*

Clients want to access their data at cloud computing storage on a remote server. Logging to the stored data using a static password will not provide a high-security level because of easiness hacking static password. In [13], the author recommends using a one-time password (OTP) method. It means when a user's logged to their account, they should enter the 4-digit pin on a login page which sent to their phone numbers have registered during upping sign account. Three most popular techniques to create the OTP password method are:

1) Time Synchronization: In this technique, both the server and client should be synchronized using synchronous time clocks; otherwise, OTP will not generate.

2) Event Synchronization: In this initial technique the counter assigned for both the client and the server, first when the client wants to login OTP will generate from the initial value of the counter then increment the counter to use for next login. On the server-side, the server will get the information that happened at the client-side, and then generate OTP from the initial value for the counter, then increment the counter value. After that, if two passwords from the client and server match, then the server will allow the client to access her/his data stored at cloud storage.

3) Asynchronous Challenge-Response Technique: In this technique, each time the server will provide a challenge to the client, which is dynamically unique every time. A hacking password will be complicated when used this technique.

C. *Software Security [14]*

A program composed by all kinds of people, and some are free that means open-source software allows both the developers and the hackers to edit the code; the hacker can exploit open-source code to find the failure points to install malicious on the cloud environment. So, security of the software can be enhanced as follows.

- *Virtualization:* Using virtualization technique provides security, this technique also allows isolation between hardware and lower-level functionality. Using this technique means the different process will run above the virtual isolation server when one of the VM is hacked, then the other VM will not be affected by this.

- *Host operating system:* This is one of the most important components of the cloud computing environment if hackers can access this OS, they can pierce all the guest operating systems on the computer. Therefor; host operating system should be secure, easy to update, and maintain.

- *Guest operating system*: Different types of operating systems can be used by clients to run their virtual private server (VPS). They can create, update, delete, and modify it. Since clients responsible for achieving security on their VMs; Awareness of customers will be very necessary, they should be informed about the importance of having the last updated version of OS and update of services and products; to avoid any security halls that can be exploited by the hackers.

- *Data encryption*: Using one of the encryption method keeping your data safe from hackers

D. *Physical Security*

A physical component in the cloud environment should be secure such as software; to avoid any vulnerability on the cloud environment can be exploited by the attackers. The enhancement of the security at the physical layer can be made by:

- *Backup*: For both customers and provider, it is essential to keep an offline backup of all their files.

- *Server location*: There are many factors that should be applied to the server location. At first server room should be isolated; it doesn't have windows and should be tight security to avoid any unauthorized access. Fire extinguishing system should be activated; also, cooling systems should be provided to prevent any overheating that may happen for machines.

- *Firewall:* Its a piece of hardware that helps screen out hackers, viruses, and worms that try to reach the computer over the Internet. Each client should be provided with the complete firewall solution by the cloud computing service provider. One of the most critical functions of the firewall is to protect against DDoS attacks.

E. *Algorithmic Approach for Securing Could*

- *Secure Data Division Algorithm*: The goal of this algorithm sharing data in a secure way, *(D)* is the whole data, where *(D1, D2, D3……, Dn)* are data pieces, *(S(D1), S(D2), …. S(Dk))* is the secure methods. This technique works as the following:

Algorithm [15]

```
S(D)
{If S(D) then return;
Else
{Divide D into smaller instanced D1, D2. ⋯. Dk, $K \geqslant 1$
Apply Secure to each of these Data {D1, D2………… Dk}
Return Combine(S(D1), S(D2), …. S(Dk));
}}
```

- *Defense System for Advanced Persistent Threat*: It has been recommended that using a layered defense solution will be more secured compared to the single defense system. The basic idea of this system is to divide the cloud environment into several layers, and each layer will be responsible for a set of function to detect and protect the system from viruses and malicious. Using layered defense will provide a comprehensive approach for security to the entire component on the cloud environment [15].

VII. Homomorphic Encryption

In general, all the data that is stored in Cloud is encrypted, if the user needs to perform any computing on data, the cloud provider will decrypt the data, and then provides the decrypted data to the user. While the user process encrypted data on Cloud, it becomes prone to hacking; therefore, a new procedure technique was found to prevent data hacking and called Homomorphic Encryption. Homomorphic Encryption is a method that allows the user to perform an operation at ciphertext [18]. In addition, when the user decrypts the ciphertext, the process that is performed on it will appear in plain text. Homomorphic Encryption grants data security when storing, transmitting, and processing as shown in Fig. 4.

Let m be a plain text.

$$\text{Operation}\ (m) = \text{decrypt}\ (\text{operation}\ (\text{encrypt}\ (m))) \tag{1}$$

The Homomorphic operations are; the addition that performs on positive real numbers R+ and multiplication of set of logarithms R*as reported in [16].

Let x, y and z belong to R+

$$\text{if}\ x.y=z \tag{2}$$

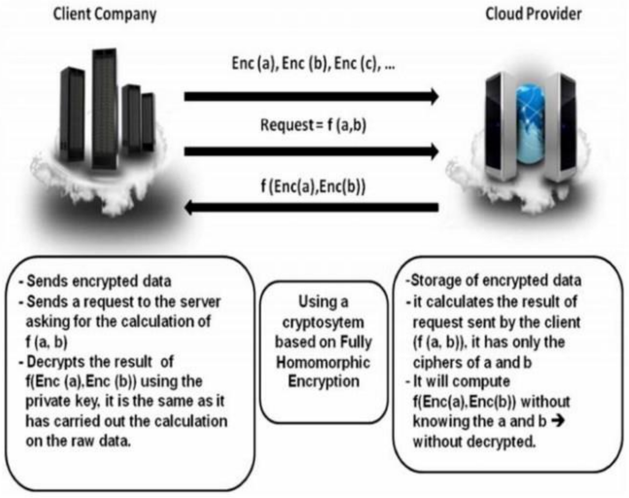
Then

$$\log(x) + \log(y) = \log(z) \tag{3}$$

Or

$$\log(x) + \log(y) = \log(x * y) \tag{4}$$

The above two formulas help us to find the value of z directly, or through logarithms. In two cases, same result can be achieved, so it's more secure to perform computation on encrypted data than decrypting it. The basic concept in Homomorphic Encryption is shown in Fig. 5. Assume that a user wants to add two numbers 10 and 15, and the two numbers are encrypted into 100 and 150, then the encrypted two numbers will be added to each other on cloud servers. When the user wants to decrypt the result, he will gain 25.



- Sends encrypted data
- Sends a request to the server asking for the calculation of f (a, b)
- Decrypts the result of f(Enc (a).Enc (b)) using the private key, it is the same as it has carried out the calculation on the raw data.

Using a cryptosytem based on Fully Homomorphic Encryption

-Storage of encrypted data
- it calculates the result of request sent by the client (f (a, b)), it has only the ciphers of a and b
- It will compute f(Enc(a).Enc(b)) without knowing the a and b ➔ without decrypted.

Fig. 4. Homomorphic Encryption Applied to Cloud Computing [17].

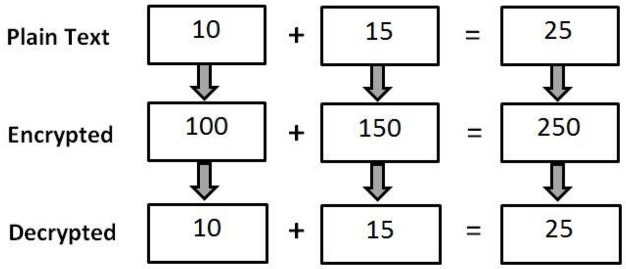


Fig. 5. Homomorphic Encryption Example [17].

There are three types of Homomorphic Encryption which can be presented as follows.

1) Fully Homomorphic Encryption (FHE), where an addition and multiplication are both performed on encrypted data with the full operation.

2) Partially Homomorphic Encryption (PHE), in this type of encryption, only one operation can be performed on encrypted data by either addition or multiplication. Pillars cryptosystem perform addition operation only while RSA cryptosystem performs multiplication operation on data.

3) Somewhat Homomorphic Encryption (SWHE), where the operation is performed on the limited number of multiplication or addition, and it is faster than FHE.

The two types (PHE and SWHE) are most widely used in cloud computing systems. Fig. 6 shows a schematic diagram of the HE system. To perform FHE on data stored in cloud servers, a secret key and a public key are needed. The following example will illustrate the FHE procedure, where J and K represent secret keys, P0, and P1 are public keys, N represents the user input. Fig. 7 describes FHE procedure. Craig Gentry from IBM has suggested the first "Full Homomorphic Encryption System" that calculates an arbitrary number of additions and multiplications, and thus computes any type of function on encrypted data. The interior working inserts another layer of encryption every step and uses an encrypted key to lock the inner layer of scrambled encrypted data. If the cipher text is decrypted, then the data will be refreshes without exposing it, allowing an infinite number of computations on it as shown in Fig. 6.

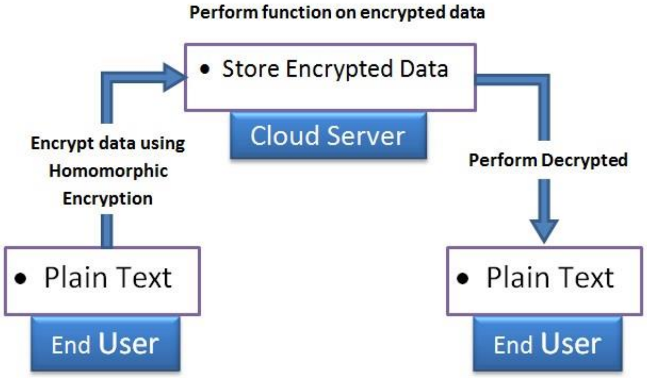


Fig. 6. FHE Proposed System [19].

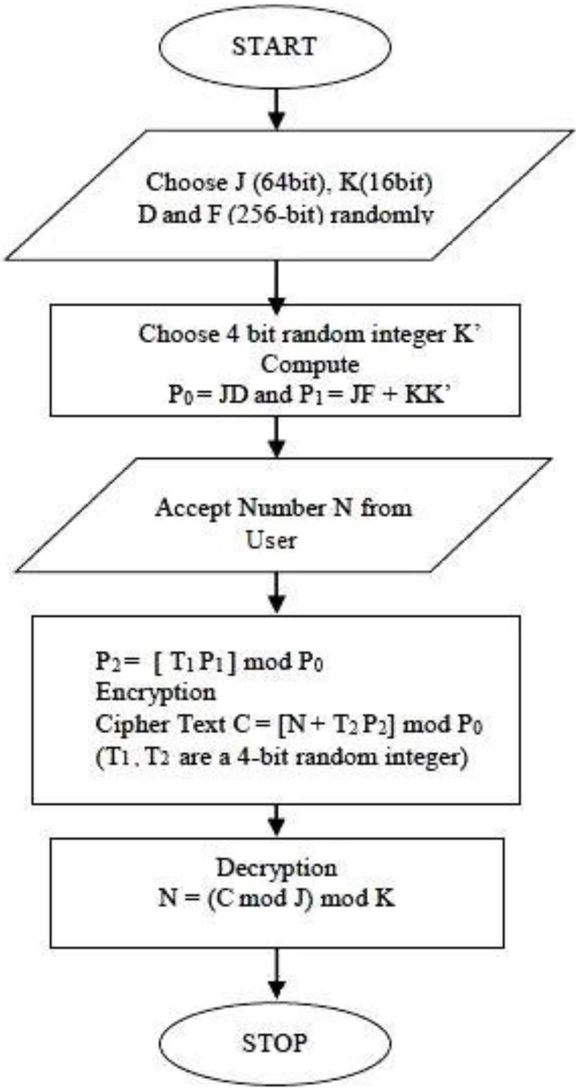

START
Choose J (64bit), K(16bit)
D and F (256-bit) randomly
Choose 4 bit random integer K'
Compute
P_0 = JD and P_1 = JF + KK'
Accept Number N from User
P_2 = [$T_1 P_1$] mod P_0
Encryption
Cipher Text C = [N + $T_2 P_2$] mod P_0
(T_1 , T_2 are a 4-bit random integer)
Decryption
N = (C mod J) mod K
STOP

Fig. 7. Flowchart of Fully Homomorphic Encryption Scheme [14].

The multi-layers of FHE cause the system to run too slowly. To solve this problem, many researchers have combined multiple schemes (see Fig. 8). The system starts with Homomorphic Encryption with a decryption algorithm embedded in a garbled circuit, which protects itself by Attribute-Based Encryption which ensures sustainable encryption [20].

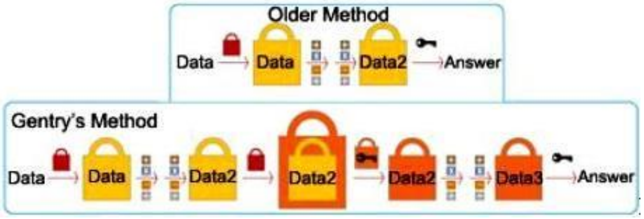


Fig. 8. Craig Gentry Implementation of FHE [20].

VIII. Conclusion

Although Cloud computing enhances the use of resources economically, there is a considerable amount of challenges it has to overcome. The main problems that arise due to the broad access nature of the networks are privacy, confidentially, and data security. Many encryption techniques are used to tackle these challenges such as Homomorphic encryption technique which is among the best encryption techniques to protect the data privacy in Cloud. It is known that all homomorphic techniques for Encryption, either partially or fully or somewhat, permit the processing of the encrypted data, which increases its security. In this review paper, some techniques and schemes for Homomorphic Encryption are discussed. In this paper, the approach of Homomorphic Encryption in cloud computing is presented and most security challenges at different levels of cloud computing with applications of vulnerabilities are explained. In addition, the most popular methods used to achieve the required level of security are presented as well. Cloud computing provides many facilities, flexibility, availability, but it faces security issues, so stringent security enforcement should be applied to ensure that the IT environments are more secure. However, it is essential to understand the security challenges and risks to avoid these challenges. On the cloud environment, all parties (customers, providers, network) should put further efforts than the traditional security solutions, because of the complex and dynamic nature of cloud computing.

Finally, several proposed solutions to migrate threats and attackers are suggested in this work as follows:

- Use more than one security layers on the application (e.g. factor authentication).
- Authentication and identity access management.
- Data encryption: when data transferred out of your network, it should be encrypted in cloud server.it is the best solution to secure information.
- Organizations should use Cloud-based security tools.
- Organizations should select an appropriate cloud model (e.g., group can use a hybrid model if they need to employ personal information on the private model, and they can use a public model to manage application).
- Do not use vendor-supplied default for any security parameters.
- Use high-security application interfaces.
- Isolation multi-tenant systems can be achieved by using isolation and segmentation techniques.
- The cloud user should have awareness of security; this creates a strong relationship between provider and customer.

References

[1] "United States : SOFTWARE ALLIANCE Hails Launch of US Framework for Improving Critical Infrastructure Cybersecurity." MENA Report, Albawaba (London) Ltd., Feb. 2014.

[2] P. Trenwith and H Venter,"A Model Aimed at Controlling the Flow of Information Across Jurisdictional Boundaries" International Conference on Cyber Warfare and Security, Academic Conferences International Limited, pp. 510, Jan. 2015,

[3] N. Dowlin, R. Gilad-Bachrach, and K. Laine,"Manual for using homomorphic encryption for bioinformatics," Proceedings of the IEEE , 2017.

[4] Top 10 Most Popular Code Review Tools for Developers https://www.softwaretestinghelp.com/code-review-tools/ (Accessed on 6th June 2019)

[5] Cloud Computing Security Benefits https://digitalguardian.com/blog/cloud-computing-security-benefits (Accessed on 1st July 2019)

[6] C. Prakash and S. Dasgupta, "Cloud computing security analysis: Challenges and possible solutions," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, pp. 54-57, 2016.

[7] K. Gupta,B. Rydhm, and B. Veerawali, "Cloud Computing: A Survey on Cloud Simulation Tools," International Journal for Innovative Research in Science and Technology, vol. 2, 2016.

[8] M. M. Alani,"Securing the Cloud: Threats, Attacks and Mitigation Techniques," Journal of Advanced Computer Science & Technology, vol. 3, 2014.

[9] V. Ashktorab, and R. T. Seyed, "Security threats and countermeasures in cloud computing," International Journal of Application or Innovation in Engineering & Management (IJAIEM) vol. 1, 2012.

[10] A. Murray, B. Geremew, N. Ebelechukwu, B. Jeremy, and P. Wayne, "Cloud Service Security & Application Vulnerabilit,." In Southeast Con, pp. 1-8. IEEE, 2015.

[11] S. Kuila, S. Shruthi, P. Chandan, and N. Ch SN Iyengar,"Cloud Computing Security by Using Mobile OTP and an Encryption Algorithm for Hospital Management," Journal of Computer and Mathematical Sciences vol. 7, 2016.

[12] E. Mathisen,"Security challenges and solutions in cloud computing," In 5th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2011).

[13] Ch. Sasanapuri, H. Chilsi, Ch. Sudhakar, and Ch. Narasimham,"Classification of APT's and Methodological Approach to Secure Cloud Services," International Journal of Applied Engineering Research, vol 11, 2016.

[14] K. K. Chauhan, A. Sanger, and A. Verma, "Homomorphic Encryption for Data Security in Cloud ," IEEE, pp. 206-209, 2015.

[15] G. Rastogi and R. Sushil,"Cloud Computing Security and Homomorphic Encryption," IUP Journal of English Studies, pp. 47-59, 2015.

[16] K. Hashizume, D. G. Rosado, E. Fernández- Medina, and E. B Fernandez, "An analysis of security issues for cloud computing," Springer, pp. 1-13, 2013.

[17] M. M. Potey, C. A. Dhoteb, and D. Sharma, "Homomorphic Encryption for Security of Cloud Data," ScienceDirect, pp. 175 – 181, 2016.

[18] S. Bajpai and P. Srivastava, "A Fully Homomorphic Encryption Implementation on Cloud Computing," International Journal of Information and computation technology, vol.4, 2014.

[19] I. Ahmad and A. Khandekar, "Homomorphic Encryption Method Applied to Cloud computing," International Journal of Information and Computation Technology, vol. 4, pp. 1519-1530, 2014.

[20] R. M. Pir, Rumel M Pir, and I. U. Ahmed, "A Survey on Homomorphic Encryption in Cloud Computing," IJEDR, vol. 2, pp. 2173-2177, 2014.

Video Analysis with Faces using Harris Detector and Correlation

Rodolfo Romero Herrera[1]
Departamento de Ciencias e Ingeniería de la Computación
Instituto Politécnico Nacional-ESCOM, Ciudad de México, México

Francisco Gallegos Funes[2]
Sección de Estudios de Posgrado e Investigación, Instituto Politécnico Nacional- ESIME, Ciudad de México, México

José Elias Romero Martínez[3]
Sección de Estudios de Posgrado e Investigación, Instituto Politécnico Nacional, Ciudad de México, México

***Abstract*—A procedure is presented to detect changes in a video sequence based on the Viola & Jones Method to obtain images of faces of persons; videos are taken of the web. The software allows to obtain images or frames separated from nose, mouth, and eyes but the case of the eyes is taken as an example. Change detection is done by using correlation and the Harris detector. The correlation results allow us to recognize changes in position in the person or movement of the camera if the individual remains fixed and vice versa. It is possible to analyze a part of the face thanks to the Harris detector; It is possible with detected reference points to recognize very small changes in the video sequence of a particular organ; such as the human eye, even when the image is of poor quality as is the case of videos downloaded from the Internet or taken with a low-resolution camera.**

Keywords—Harris detect; Viola and Jones; Harris detector; correlation; video

I. Introduction

Face analysis is a method used in several applications such as cell phone security systems, to detect stress levels, in gender recognition, etc. [1] [2] [3]; Another topic of interest is undoubtedly the follow-up of the face and its analysis in video sequences [4] [5]; and it is because there is more information in a temporary space than if only one image is analyzed; since, without a doubt, the dependence between one image and another on a video sequence reveals information of interest; However, the problem is complicated, if the video is of poor quality as are internet videos.

The relationship between frames or consecutive images can give us, for example, the recognition of an important change pattern such as the state of emotion between sequences or a significant change due to an unexpected event [6] [7]. This can be measured by the correlation in contours or landmark [8] [9]. For example, if you want to measure the duration of an emotion or the unexpected change of it [10]. Thus applying some processing to a bad video will result in a bad analysis. For this reason, it is proposed to apply such analysis to landmarks, in order to obtain results for these cases, no matter the poor of the video.

In this investigation, the Viola & Jones method is used for the location of the face in video sequences and the segmentation in RGB for its follow-up [11] [12]. Subsequently, a probabilistic analysis is used to see the relationship between one image and another and detect changes [13]. It is feasible to separate parts of the face in the mouth, nose, and eyes, for the analysis of their individual sequence. It was chosen to process the eyes because it involves greater difficulties; because processing smaller images or frames is more difficult, Due to the fact, they are taken from bodies or faces whose quality is low. Thus techniques such as segmentation or mathematical morphology are complicated yield poor results.

Processing is important as it allows the study of recognition of temporal space patterns in people's faces or individual analysis of the characteristics of the eyes, nose, and mouth. With the applied techniques it is feasible to analyze the behavior of a part of the face without having expensive cameras or search for people on video from the internet.

II. Viola and Jones Method

A. Method

Viola & Jones classifies the image by value with simple characteristics of three types; square entities with sub-boxes called rectangular components [14]. See Fig. 1. These components are shown in Fig. 1, where the grayscale image is scanned by each component to look for positive features with AdaBoost and cascading clarifiers. When a face is detected, a rectangular contour is drawn around the face. The value of the characteristic is the difference between the black and white regions [15]. However, the total number of Haar features is very large, much compared to the number of pixels; AdaBoost is used to select the specific Haar function used and set the threshold value. To ensure rapid classification, the learning process must eliminate most of the available functions and focus on a small set of features [16].

A classification method considers levels of selection. Each level has been trained using the Haar function. The selection separates sub-windows that contain positive objects (images with the desired object) and negative objects (images that do not have the desired object). The Viola-Jones method combined four general rules: Haar Feature, Integral Image, Adaboost learning, and Cascade classifier.

Haar Feature values are obtained from the difference between the numbers of dark areas pixel values minus the number of bright area pixels:

$$\mathrm{F(Haar)} = \sum \mathrm{F}_{\mathrm{white}} - \sum \mathrm{F}_{\mathrm{black}} \qquad (1)$$

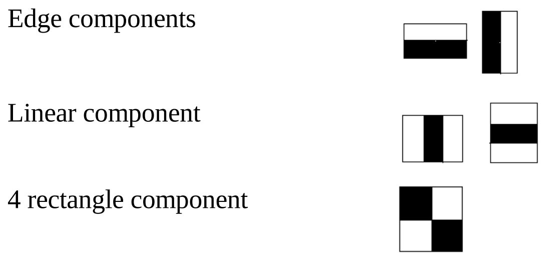


Fig. 1. Haar Rectangular Components.

Where $\sum F_{white}$ is the characteristic value in the brightest area and $\sum F_{Black}$ is the value in the dark area. Haar features are made up of two or three rectangles. The images are scanned and the Haar characteristics in the current stage are searched. The weight and size of each function and the functions themselves are generated using an AdaBoost algorithm [17]. The integral image generated is a technique to quickly calculate the value of the characteristic, and change the value of each pixel to another image representation. See Fig. 2. The integral image in ii(x,y) can be found by equation (2).

$$ii(x,y) = \sum x^{'} \leq x, y^{'} \leq \; y^{i(x^{'},y^{'})} \qquad (2)$$

where ii (x,y) is the integral image at (x,y) and i (x',y') is the pixel value of the original image.

For the overall image in general, small units are added simultaneously, in this case, are pixels. The integral value for each pixel was the sum of all pixels from top to bottom. Starting from the upper left to the lower right, the entire image is the sum with multiple integer operations per pixel. The value of a characteristic is calculated with the value of the integral image at four points. See Fig. 3. If the integral value of the image of point 1 is A, point 2 is A + B, point 3 is A + C, and at point 4 it is A + B + C + D, then the number of pixels in region D is calculated by points 4 + 1 (2 + 3).

The combination of weak classifiers is used to improve the classification. A cascade classifier is a combination of classifiers with a multilevel structure that increases the speed of object detection by focusing only on image areas. See Fig. 4.

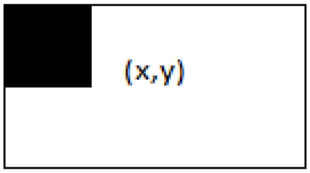


Fig. 2. Integral Image (x, y).

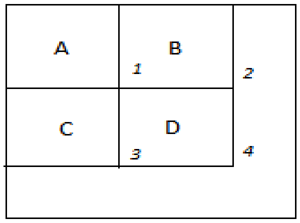


Fig. 3. The Score Count of the Figure.

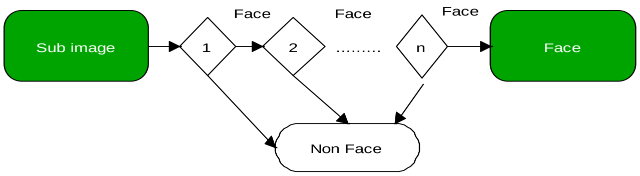


Fig. 4. Cascade Classifier.

A weak classifier is defined by equation (3):

$$h_j(x) = \begin{cases} 1, & if\ p_j f_j \ < \ p_j\ \theta_j(x) \\ 0, & \qquad other \end{cases} \qquad (3)$$

Where *ℏj (x)* is a weak classification, p_j is even to j, Θ_j is the threshold to j and x is a dimension of the second image. So the strong classifier is:

$$h(x) = \begin{cases} 1, \sum_{t=1}^{T} \alpha_t\, h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0, \qquad\qquad other \end{cases} \ where\ \alpha_t = \log \frac{1}{\beta_t} \ (4)$$

Facial detection using the Viola & Jones method is presented in Fig. 5. First, the image that can contain a face is read. Second, the Haar feature is processed in the images, which results in the difference of the threshold value of the dark areas and bright areas of the image. If the difference between dark and bright areas is above the threshold value, then there is a face within the image. Face detection is achieved by combining some AdaBoost classifiers as efficient filters that classify the image area. If any of the filters prevents the image from passing, then it has no face. However, when passed through all filters, the image area was classified as a face. The order of the filters in the cascade is determined by the weight given by AdaBoost. The largest weighted filter is placed in the first step, in order to ignore deleting the faceless image area as soon as possible.

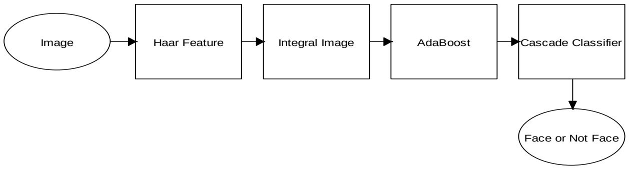


Fig. 5. Face Detection Process with the Viola-Jones Method.

III. Image or Frame Generator System Design

A. Processing

To carry out the processing of the body parts, we must first have the object to be processed, for which a previously stored video file can be opened or a camera that allows the capture of the sequence of frames can also be used. Due to the existence of video files of different dimensions, it is necessary to resize it to observe properly in the user interface. See Fig. 6.

Images can be generated from a video or directly from a webcam, which is useful for comparisons or analysis. See Fig. 7.

The results improve if the face is first located and on that plane, the mouth or eyes are located. Face localization is performed using the Viola & Jones method. It was conceivable to generate images of each of the parts of the face and perform the probabilistic analysis.

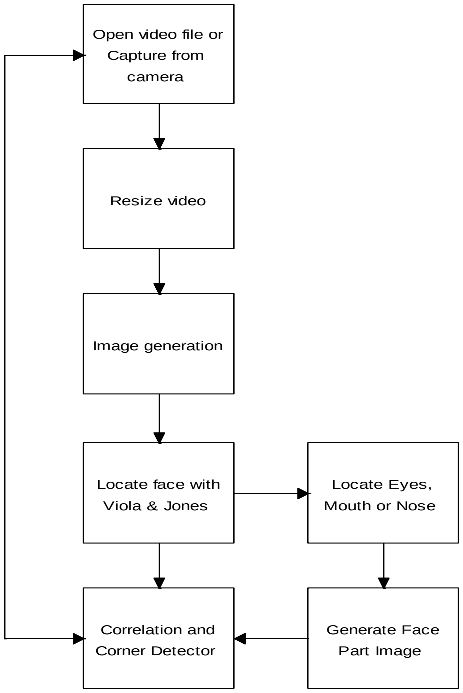

Open video file or Capture from camera
Resize video
Image generation
Locate face with Viola & Jones
Locate Eyes, Mouth or Nose
Correlation and Corner Detector
Generate Face Part Image

Fig. 6. Procedure for Face Analysis.

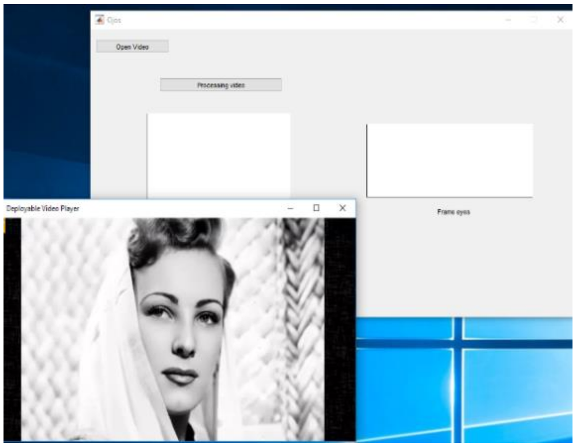


Fig. 7. User Interface “Open Video File”.

Fig. 8 shows the results of locating the eyes of the original video. The high-definition and closeness of the camera to the face make it easy to get good results.

Fig. 9 shows an image of a video with low resolution, full bode and away from the camera; which is very common in internet videos; still, the eyes are located with the method of Viola & Jones; but techniques such as contour detect or segmentation yield poor results. Although, the eyes segmentation is more successful than detector of contour Sobel; however, a histogram threshold is required for each frame or image; it is clear that methods must be rejected and opt for others to detect movement; since the problem can be solved with high resolution cameras as in [18] which increases the cost. Another alternative is the use of specialized hardware based on FPGA, but most people do not have such an option [19]. For this reason, it was decided to solve the problem by means of correlation and algorithm of Harris.

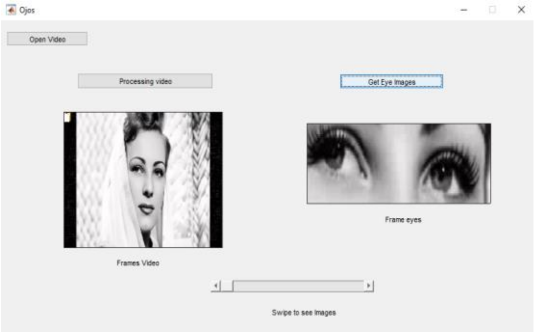
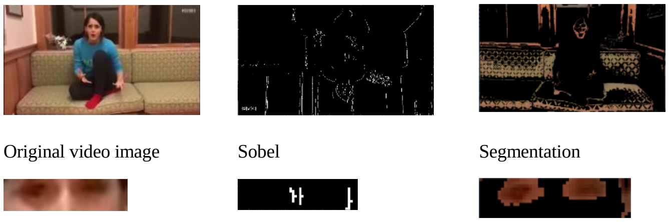


Fig. 8. Obtaining Images.



Original video image

Sobel

Segmentation

Image of eyes after Viola & Jones

Sobel

Segmentation

Fig. 9. Images using Sobel Detector and Segmentation.

IV. Analysis of Correlation and Harris using

To complete the project, an analysis was performed by using the correlation between images and the Harris detector The correlation coefficient is calculated by equation (5) [20].

$$r = \frac{\sum_m \sum_n (A_{mn} - \overline{A})\,(B_{mn} - \overline{B})}{\sqrt{\sum_m \sum_n (A_{mn} - \overline{A})^2\,(\sum_m \sum_n (B_{mn} - \overline{B})^2}} \qquad (5)$$

Where the means are is equal to $\overline{A} = mean2(A)$, y $\overline{B} = mean2(B)$.

A. Harris-Laplaciano Corner Spot Detector

The corner points in an image are the points that have a significant intensity gradient in the cardinal axes. Harris corners have invariability to rotation and gray level change [21]. Harris's algorithm detects corner points.

It is required to estimate the proper value of the Harris matrix. The Harris matrix is a symmetric matrix similar to a covariance matrix. The main diagonal is composed of the two averages of the square gradients. The elements outside the diagonal are the averages of the cross product of the gradient $\langle G_{xy} \rangle$. Matrix of Harris is:

$$A_{Harris} = \begin{bmatrix} \langle G_x^2 \rangle & \langle G_{xy} \rangle \\ \langle G_{xy} \rangle & \langle G_y^2 \rangle \end{bmatrix} \qquad (6)$$

Consider first the measure of the corner response R. The contours of the constant R are shown by thin lines. R is positive in the corner region, negative in the border regions and small in the flat region. Increasing the contrast increases the magnitude of the response. The flat region is specified by T_r, which falls below some selected threshold.

The key simplification of the Harris algorithm is to estimate the proper values of the Harris matrix as a determinant minus the scaled trace or flat squared region.

$$R = \det(A_{Harris}) - k\, T_r^2(A_{Harris)}) \qquad (7)$$

Where k is a constant typically with a value of 0.04. R can also be expressed with gradients:

$$R = \left(\langle G_x^2\rangle\langle G_y^2\rangle - \langle G_{xy}\rangle^2\right) - k\left(\langle G_x^2\rangle + \langle G_y^2\rangle\right)^2 \qquad (8)$$

So that when the response is greater than a predefined threshold, a corner is detected:

$$R > k_{thresh}$$

$$\left(\langle G_x^2\rangle\langle G_y^2\rangle - \langle G_{xy}\rangle^2\right) - k\left(\langle G_x^2\rangle + \langle G_y^2\rangle\right)^2 > k_{thresh} \qquad (9)$$

With this, a pixel in the corner region (positive response) is selected if your response is a local maximum of 8 ways. Similarly, the pixels of the border region are considered edges if their responses are local and negative minimums in the x or y directions, depending on whether the magnitude of the first gradient in the x or y direction, respectively, is greater. This results in thin edges.

By applying low and high thresholds, the hysteresis of the edges can be carried out and this can improve the continuity of the edges. The processing removes stretch marks from the edges and short and isolated edges and joins brief breaks at the edges. This results in continuous fine edges that generally end in the regions of the corners. Edge terminators are then linked to corner pixels that reside within regions, to form a connected edge-vertex graph.

Algorithm of Harris uses equation (7) as a metric, avoiding any division or square root operation. Another way of doing corner detection is to calculate the actual eigenvalues.

The analytical solution for the eigenvalues of a 2x2 matrix can be used in corner detection. When the eigenvalues are positive and large on the same scale, a corner is found. In such a way that:

$$\lambda_1 = \frac{T_r(A)}{2} + \sqrt{\frac{T_r^2(A)}{4}\det(A)} \qquad (10)$$

$$\lambda_2 = \frac{T_r(A)}{2} + \sqrt{\frac{T_r^2(A)}{4}\ \det(A)} \qquad (11)$$

Substituting the gradients:

$$\lambda_1 = \left(\frac{\langle G_x^2\rangle + \langle G_y^2\rangle}{2}\right) + \sqrt{\left(\frac{\langle G_x^2\rangle + \langle G_y^2\rangle}{2}\right)^2} - \left(\langle G_x^2\rangle\langle G_y^2\rangle - \langle G_{xy}\rangle^2\right) \quad (12)$$

$$\lambda_2 = \left(\frac{\langle G_x^2\rangle + \langle G_y^2\rangle}{2}\right) + \sqrt{\left(\frac{\langle G_x^2\rangle + \langle G_y^2\rangle}{2}\right)^2} - \left(\langle G_x^2\rangle\langle G_y^2\rangle - \langle G_{xy}\rangle^2\right) \quad (13)$$

V. Results

Tests were conducted with 60 videos taken from the internet, resulting in the detection of movement specifically in faces, eyes, and mouths.

Fig. 10 shows the correlation between consecutive images in case of the complete video. The similarity of the images is observed, due to the value close to 1 of the correlation coefficient; however, abrupt changes can be observed in frame 16 and between frames 53 and 56; caused by camera movements. So we can assume that little sensitive to this event. The statistical data obtained are shown in Table I. This table shows the correlation between consecutive images which proves that it is the same person since the values are positive and close to 1. It is important to check for some cases that it is the same person, since in the videos downloaded from the internet or taken with webcam usually changes the scenario constantly, as well as people recorded or captured on video.

The correlation between the first image and the rest of the images or frames can be made and the result is shown in Fig. 11. The graph shows greater differences, which is corroborated by Table II. This is due to the movement of the person. The range of 0.017 can be seen in the graph of Fig. 11. Although this interval is small, a change can be observed.

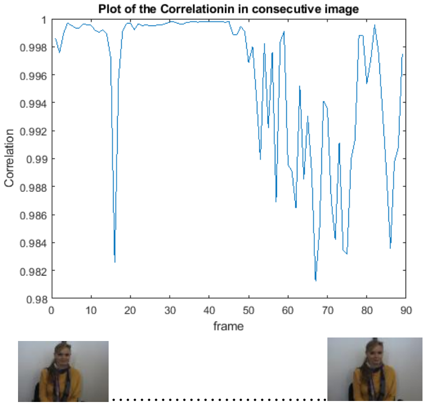


Fig. 10. Consecutive Image Correlation.

TABLE. I. Correlation Statistics between Consecutive Images

	Frame	Correlación
min	1	0.9812
max	89	0.998
mean	45	0.9958
median	45	0.9989
mode	1	0.9812
std	25.84	0.005309
range	88	0.0186

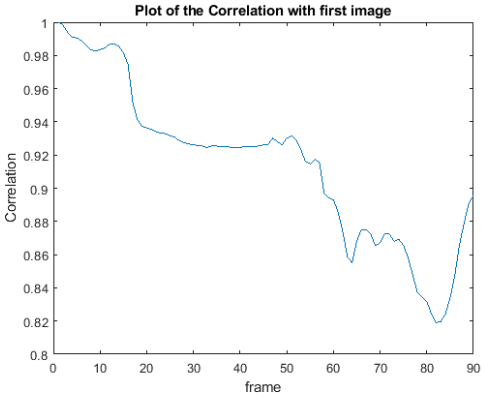


Fig. 11. Correlation between Frame 0 and the Rest of the Images.

TABLE. II. STATISTICS PRODUCED BY CORRELATION

	Frame	Correlation
min	1	0.8189
max	89	0.9986
mean	45	0.9132
median	45	0.8189
mode	1	0.8189
std	25.84	0.04722
range	88	0.01796

The analysis so far considered the total image; now a body part is taken into account. For example, the eyes of the face. For which the Viola & Jones method already described was used and the correlation was applied. Fig. 12 shows the correlation between consecutive images of the eye. It is observed that there are many points that you have no relation some; This is because in the video people open or closes their eyes constantly; for example to blink.

When the images are high resolution or the camera is approached with optical zoom to the human eye, it is possible by morphological operations to obtain the dilation of the iris and thus determine emotional states or some other statistical parameter [22]. However, when you use a webcam or video from somewhere on the Internet, it becomes a more complex process; For this reason, it is better to use corner detectors such as Harris and work with the landmark during the sequence of the event [23][24].

The Harris corner detector was applied to obtain the graph of Fig. 13. Where changes in the number of Landmarks detected are observed; which indicates that there are variations in the different frames processed.

Table III shows the statics of the detected landmarks. There is a minimum of 24 and a maximum of 42 landmarks resulting in a range of 18; which tells us the existence of differences between the number of points detected in these frames; that indicate a change due to the eye movement and that can be corroborated by looking at two images shown in Fig. 13, below the graph.

Without decreasing the number of marks you can also locate the center of the pupils and measure the separation in pixels between the eye. See Fig. 14.

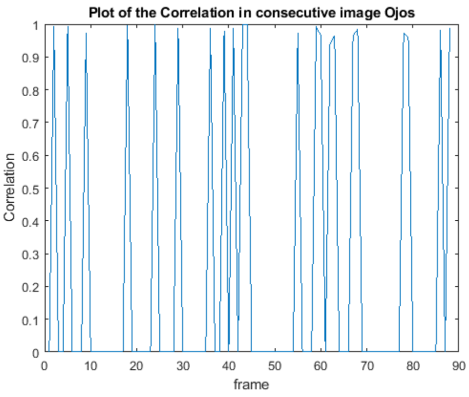


Fig. 12. Consecutive Eye Image Correlation.

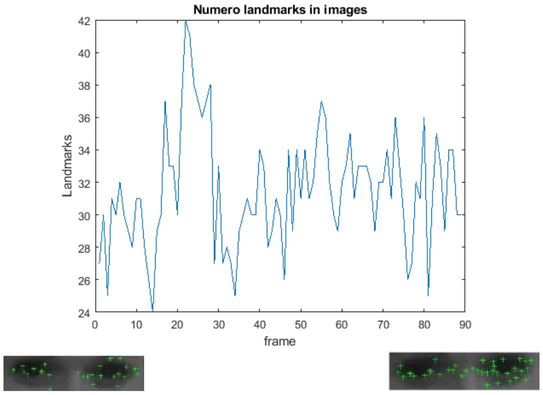


Fig. 13. Landmark Detector by Harris.

TABLE. III. LANDMARKS STATISTICS

	Frame	Landmarks
min	1	24
max	89	42
mean	45	31.51
median	45	31
mode	1	30
std	25.84	3.581
range	88	18

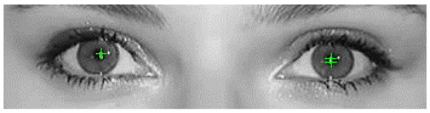


Fig. 14. Separation in Pixels between the Eyes.

VI. CONCLUSION

Movements or changes in the eyes within the image can be identified by correlation in frames or images; however. It is necessary to use methods to detect a change in the images; Changes can be recognized due to eyes movement in correlation graphs. However, the results are better with method of Harris, because the range is greater than that of the correlation.

In the case of the analysis of the eyes, changes were recorded mainly due to the blink of the person; which allows us to conclude that is possible that identify changes in the eyes of the same, even if the video is not high definition or close to the face.

The methods of Viola & Jones and Harris are known and there therefore proven; is for this reason that they are employees. Its use allows motion detection even when you do not have High Definition or hardware at a low cost; since only one web cam is needed, and also is had the advantage of being able to process internet video.

VII. FUTURE WORK

Obtained landmarks by Harris detector, you can use other detectors or methods such as HOG and BRISK, and determine the amount of movement to relate it to affective states or specifically to stress.

ACKNOWLEDGMENT

The authors appreciate the support received by Instituto Politécnico Nacional and Consejo Nacional de Ciencia y Tecnologia.

REFERENCES

[1] D. J. Robertson, R. S. S. Kramer and A. M. Burton, "Face averages enhance user recognition for smartphone security," Plos One, vol. 10, (3), pp. e0119460-e0119460, 2015

[2] T. Chen et al, "Detection of Psychological Stress Using a Hyperspectral Imaging Technique," IEEE Transactions on Affective Computing, vol. 5, (4), pp. 391-405, 2014.

[3] J. Bekios Calfa, J. M. Buenaposada and L. Baumela, "Class–Conditional Probabilistic Principal Component Analysis: Application to Gender Recognition," Class–Conditional Probabilistic Principal Component Analysis: Application to Gender Recognition, 2011.

[4] D. O. Gorodnichy, "Seeing faces in video by computers. Editorial for Special Issue on Face Processing in Video Sequences," Image and Vision Computing, vol. 24, (6), pp. 551-556, 2006.

[5] Zhengrong Yao, Haibo Li, Tracking a detected face with dynamic programming, 1 June 2006, pp.

[6] B. App, C. L. Reed and D. N. McIntosh, "Relative contributions of face and body configurations: Perceiving emotional state and motion intention," Cognition & Emotion, vol. 26, (4), pp. 690-698, 2012.

[7] L. A. Stockdale et al, "Emotionally anesthetized: Media violence induces neural changes during emotional face processing," Social Cognitive and Affective Neuroscience, vol. 10, (10), pp. 1373-1382, 2015.

[8] R. Gonzalez and R. Woods, Digital image processing. New Delhi: Dorling Kindersley, 2014.

[9] K. Rohr and SpringerLink (Online service), Landmark-Based Image Analysis: Using Geometric and Intensity Models. 200121. DOI: 10.1007/978-94-015-9787-6.

[10] M. Codispoti, M. Mazzetti and M. M. Bradley, "Unmasking emotion: Exposure duration and emotional engagement," Psychophysiology, vol. 46, (4), pp. 731-738, 2009.

[11] E. Winarno et al, "Multi-view faces detection using viola-jones method," in 2018, . DOI: 10.1088/1742- 6596/1114/1/012068.

[12] Gonzalo Pajares, Jesús M. de la Cruz, "Visión por computadora, Imágenes Digitales y apliaciones", Alfaomega, México 2002.

[13] E. Cuevas, D. Zaldívar and M. Pérez-Cisneros, Procesamiento digital de imágenes usando MatLAB & Simulink. México, D.F, 2010.

[14] Damanik, Rudolfo Rizki, et al. "An application of viola jones method for face recognition for absence process efficiency." Journal of Physics: Conference Series. Vol. 1007. No. 1. IOP Publishing, 2018.

[15] E. Winarno et al, "Multi-view faces detection using viola-jones method," in 2018, . DOI: 10.1088/1742- 6596/1114/1/012068.

[16] Haralick, Robert M., and Linda G. Shapiro, Computer and Robot Vision, Volume II, Addison-Wesley, 1992, pp. 316-317.

[17] Y. Li, W. Shi and A. Liu, "A Harris Corner Detection Algorithm for Multispectral Images Based on the Correlation," IET Conference Proceedings, 2015.

[18] Smith, M., Maiti, A., Maxwell, A.D., Kist, Colour Histogram Segmentation for Object Tracking in Remote Laboratory Environments A.A. (2020) Lecture Notes in Networks and Systems, 80, pp. 544-563.

[19] Liu, B. Real-Time Video Edge Enhancement IP Core Based on FPGA and Sobel Operator (2020) Advances in Intelligent Systems and Computing, 928, pp. 123-129.

[20] Lewis, J. P., "Fast Normalized Cross-Correlation," Industrial Light & Magic.

[21] D. Yang et al, "A method to detect landmark pairs accurately between intra-patient volumetric medical images," Medical Physics, vol. 44, (11), pp. 5859-5872, 2017.

[22] Herrera, Rodolfo Romero, Francisco Gallegos Funes, and Saul De La O. Torres. "Graphing emotional patterns by dilation of the iris in video sequences." IJACSA Editorial (2011).

[23] Harris, C; and M.J. Stephens. " A combined Corner Edge Detector", Proceedings of the 4th Alvey Vision Conference. Agust 1988, pp. 147-152.

[24] Mikolajczyk, K., Schmid, C. A performance evaluation of local descriptors (2005) IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (10), pp. 1615-1630.

Mapping of Independent Tasks in the Cloud Computing Environment

Biswajit Nayak[1], Sanjay Kumar Padhi[2]
Computer Sc. & Engineering
Biju Patnaik Technical University
Rourkela, Odisha, India

***Abstract*—Cloud computing is a technology that provides many resources and facility to share data. Due to the concept of open environment in the cloud computing the request or data increases quickly. So this problem can be solved by proper utilization of tasks along with available resources. Task scheduling algorithm plays an immense role in the cloud computing environment in minimizing the time required for completion of the task assigned to the resource available. There are several algorithms introduced to solve the problem of scheduling task of several kinds but all the developed algorithms are task dependent algorithms. The major criteria of the task scheduling algorithm are to optimize resource utilization in the diverse computing environment, so as to minimize makespan and execution time so that the accountability of healthcare industry that uses cloud computing can be enhanced. The proposed algorithm is designed to deal with variable length tasks by taking the advantages of the different heuristic algorithm and ensures optimum task scheduling with various available resources to enhance the quality of the healthcare system.**

Keywords—Scheduling; mixed model; cloud computing; makespan; healthcare

I. Introduction

Cloud computing environment differs from traditional computing environment on the basis of the target of scheduling. In the case of traditional computing, the transferred data is small but in the case of cloud computing, the transferred data is very large. Scheduling of resources ensures better service without any interrupt. This technique not only manages of task load but also fulfil the requirement of allocation of task dynamically with availability, flexibility, minimal cost and scalability features. The load balance technique ensures the availability of resource on demand, Proper utilization of resources under different conditions, reduced cost of resource use, Manipulation of energy at different load conditions [1][2].

Scheduling is a mechanism to maximize the throughput, required utilization of resource and also system performance through the allocation of task or job to the resource available. Due to increase in demand for technology with minimal cost and quick access time, some task scheduling prototype required. The processes start as the user submit tasks to the scheduler.

The scheduler schedules the task according to the availability of resources. As Fig. 1 shows in the job allocation process the scheduler allocates the job based on the cloud information repository. The datacenter computes the job within the stipulated time period. The task or job may be considered as data insertion, processing or accessing inserted data, software or it may be storage functions [3][4][5].

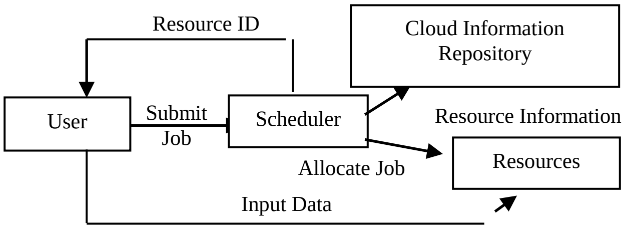


Fig. 1. Job Allocation Process.

II. Scheduling Parameters

There are certain cloud-computing performance metrics that are responsible for effective load balancing [6] [7][8]:

Throughput (TP): It is calculated to determine the performance of the system by calculating the number of tasks executed in one-unit time. It is measured by comparing with makespan. Increase in makespan reduces the performance and also decrease in makespan means optimum throughput.

Thrashing (TH): Thrashing takes place due to memory and other limited or exhausted resources. This occurs due to improper schedule of tasks, so good algorithm is required to maintain available resources.

Reliability (R): A system must be reliable to gain the faith of the user. If a task is transferred to any other virtual machine due to failure in task execution, then it can be treated as reliability of system. In other terms, a system is reliable if the system will work efficiently even if system fails to execute some of the task dedicated to.

Accuracy (A): This parameter ensures whether the result of the execution of the task meeting to the required result or not. If it will match the result, then it is accurate otherwise not.

Predictability (PR): The system should have the capability to predict the task allocation, execution and time required to complete considering available resources. It is also termed as a degree of prediction. It enhances the makespan of system.

Makespan (MS): It is defined as the time required completing all the tasks. In other words, it can be defined as the maximum time required by the system running over the data centre. Makespan is directly proportional to load balance;

means, if the makespan is less the load balancing, is good. One of the major characteristics of good task scheduling algorithm is diminishing makespan.

Scalability(S): It is a concept a system that ensures the execution of tasks under different conditional environment where the number of tasks may increase or decreases unexpectedly or periodically.

Fault Tolerance (FT): It is a method that increases the performance of a system by providing uninterrupted services even if one or more elements of the system fail to work properly. It is also responsible for resolving elements of logical errors.

Associated Overhead (AO): Associated overhead is directly proportional to load balancing. If the associated overhead is more that means the load balance is not proper. If the associated overhead is less that means the load of the system is proper.

Migration time (MT): It is a time required when a task is shifted from one resource to another or from one resource to different virtual machine or migration of service from one virtual machine to another virtual machine. The larger number of migration of the virtual machine leads to poor performance of system because it degrades the makespan.

Response time (RT): Time required acknowledging task for execution is known as response time. Lesser the response time leads to the greater performance of the system.

Energy Consumption (EC): It one of the major metrics in the cloud computing environment like makespan. Energy is calculated on the basis of energy consumed by all the devices connected to the system. The devices may be- output devices, device connectors, and application servers, etc.

Resource Utilization (RU): It is a concept defines the degree of use of resources in the system. If the load balancing is maximum that means resource utilization is maximized. Load balancing is directly proportional to resource utilization.

A good technique requires a good scheduler. Let's consider there are n numbers inputs for which N numbers of VMs are available. The set of tasks is (T1, T2, T3, T4... Tn). The heterogeneous environment in cloud computing uses expected time to compute matrix for load balancing. So the value of the matrix needs to be determined.

$$T_{EC_{ij}} = L_i/P_j$$

Where: L_i – Lenth of the ith task (MI)

P_j – The processing speed of the jth virtual machine (MIPS).

The most used performance parameter is the makespan in cloud computing. The different virtual machine takes different time for execution in the cloud computing environment. Good load balancing means minimal makespan.

The execution time of jth $VM(ET)_j$ is based on the decision variableX_{ij},

where

$$X_{ij} = \begin{cases} 1 \ \ If\ T_i \ \in VM_j \\ 2 \ \ If\ T_i \ \notin VM_j \end{cases}$$

The execution time of virtual machine literally depends upon the decision variable X_{ij}.

$$T_{E\,j} = \sum_{i=1}^{n} X_{ij} * T_{EC_{ij}}$$

Makespan can be defined as the maximum time consumed by any virtual machine. So the makespan can be calculated as:

$$MS = Max[T_{E_j}]\,\begin{matrix} N \\ j=1 \end{matrix}$$

III. Basic Task Scheduling Algorithms

Completion time is the basic criteria for MinMin algorithm. It uses two different time quantum-like Execution Time and completion time. Initially, it calculates the completion time, on the basis of minimum completion it finds the task and assigns to the corresponding resource. Then the task is removed and the completion time is updated. There is no longer waiting of processor for smaller tasks but it signifies the starvation for larger tasks and failed to perform well when small tasks are more as compare to large tasks. The Min-Min heuristics ensures the completion of task execution with minimum time period as compared to the other task and allocated to the suitable machine. According to the process the small tasks are assigned first then large tasks hence the makespan increase as the completion-time increases [9] [10].

Completion time is the basic criteria of the MaxMin algorithm. It uses two different time quantum-like Execution Time and completion time. Initially, it calculates the completion time, on the basis of minimum completion it finds the tasks and assigns to the corresponding resource on the basis of maximum completion time. There is no longer waiting of processor for larger tasks but it signifies the starvation for smaller tasks and failed to perform when the large task is more than a small task. So it eliminates the problem resides in MINMIN algorithm. Like the above algorithms, some other algorithms provide same result or poor with large search space. Some of the algorithms also tried to improve the makespan and throughput performance [11] [12] [13].

IV. Proposed Algorithm

The proposed scheduling algorithm tried to eradicate the problem persists in the basic algorithm developed. Even if the complexity of First-Come-First-Serve algorithm is very less, it performed less because it used arrival time to calculate time required to complete the task. Similarly, the Round Robin algorithm used arrival time along with the time quantum to reduce the time required to complete the task but still did not perform well. Apart from the basic algorithms some heuristics are used to enhance the performance. Some load balancing algorithm used to maximize the performance but due to the simultaneous use of resources or machines the performed poor makespan. Few more algorithms like Minimum Completion Time, Min-Min, Max-Min, suffrage algorithm etc. are used to solve the problem but still lacking to provide the optimum solution [14] [15] [16].

The proposed model is a Mixed Model to solve the starvation problem present in the model discusses above.

```
Step 1:
    1. Start
    2. Compute the completion time matrix of resources
and tasks.
            For all task Ti
                For all resources Rj
                    CTij= ETij + rj
                End For
            End For
Step 2:
    3. Find the number of smallest number task "S" and
largest number tasks "L"
    4. If S>L
            Go to Step 4
       Else
            Go to Step 3
Step 3:
    5. For each task in the matrix, find the task ti
            with a minimum completion time
            and the resource on which it is
            calculated.
            Assign ti to resource Rj that has
                minimum completion time.
            Remove task ti from the matrix
            Update resource Rj ready time (rj)
            Update completion time of all un-
                mapped tasks in the matrix.
            Repeat all the steps of step3 until all the
                tasks in the matrix have been
                mapped.
       End For
    6. Go to Step 5
Step 4:
    7. For each task in the matrix, find the task ti
            with a minimum completion time
            and the resource on which it is
            calculated.
            Assign ti to resource Rj that has
                maximum completion time from
                selected minimum completion
                time.
            Remove task ti from the matrix
            Update resource Rj ready time (rj)
            Update completion time of all un-
                mapped tasks in the matrix.

            Repeat all the steps of step4 until all the
                tasks in the matrix have been
                mapped.
       End For
    8. Go to Step 5
Step 5:
    9. Display Result
    10. Stop.
```

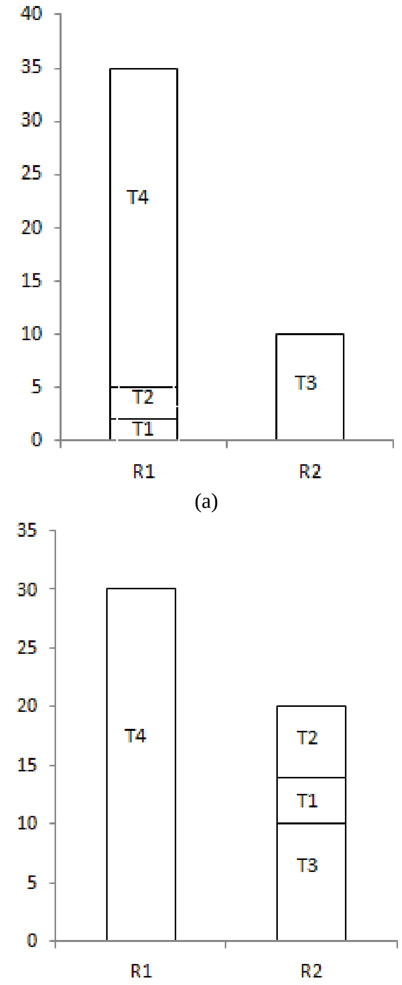
V. Performance Analysis

The proposed algorithm calculates the completion time of each task in a different machine and based on the expected completion time assign the tasks to the appropriate available resources. Let's consider four tasks (T1, T2, T3, T4) with execution time and two available resources (Table I). The table below clearly shows that the table consists of a large number of smaller tasks and a smaller number of large tasks.

In Fig. 2(a) all tasks in are executed according to their minimum completion gives a makespan of 35 whereas the Fig. 2(b) executes or sort all the tasks according to their maximum completion time and give a makespan of 30.

TABLE. I. Resources for Allocation

Resource Task	R1	R2
T1	2	4
T2	3	6
T3	4	10
T4	30	70



(a)

(b)

Fig. 2. (a) Output using STEP-3; (b) Output using STEP-4 NB: - X-axis showing the resources or machine (R1, R2) and Y-axis showing completion time (T1, T2, T3, T4).

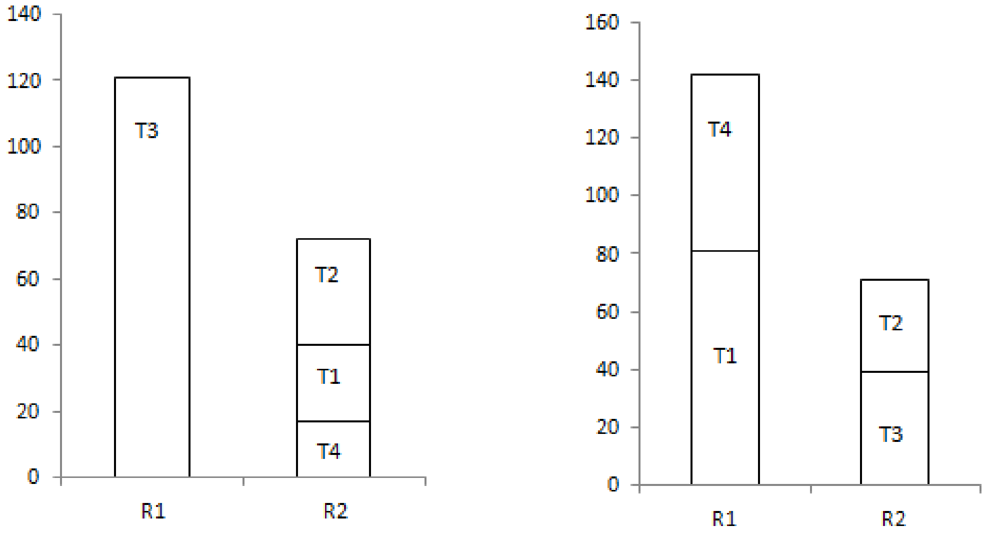
Let’s consider some inputs just opposite to the previous inputs and analyze the output to ensure that the algorithm performs better in all condition. Table II clearly shows that the table consists of a large number of large tasks and a smaller number of small tasks.

In Fig. 3, all tasks are executed according to their minimum completion gives a makespan of 121 whereas Fig.4 executes all the tasks according to their maximum completion time and give a makespan of 142. The example used consists of a large number of large tasks as compared to a number of small tasks. So it makes sure the execution of Step 3 of the algorithm and gives better makespan. Fig. 3 clearly shows the difference between the output of execution for Step 3 and Step 4. The makespan of Step3 is less as compare to the makespan of Step 4.

TABLE. II. RESOURCES FOR ALLOCATION

Resource / Task	R1	R2
T1	81	23
T2	112	32
T3	121	39
T4	61	17

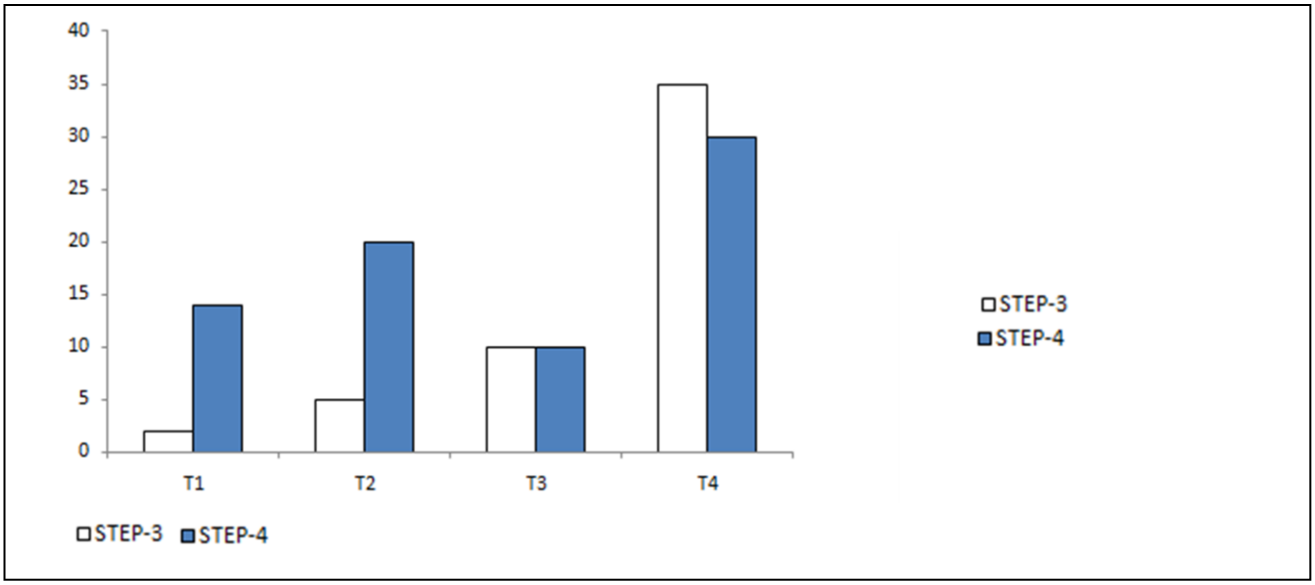
On the basis of the above analysis for the better visibility Fig. 4(a) and Fig. 4(b) shows the output by comparing the time required for the execution of tasks at different conditions.



(a)

(b)

Fig. 3. (a) Output using STEP-3; (b) Output using STEP-4 NB: - X-Axis Showing the Resources or Machine (R1, R2) and Y-Axis Showing Completion Time (T1, T2, T3, T4).



(a) Execution Time Required at different Conditions NB: - X-Axis Showing the different Task Assigned and Y-Axis Showing Completion Time Required by different Tasks.

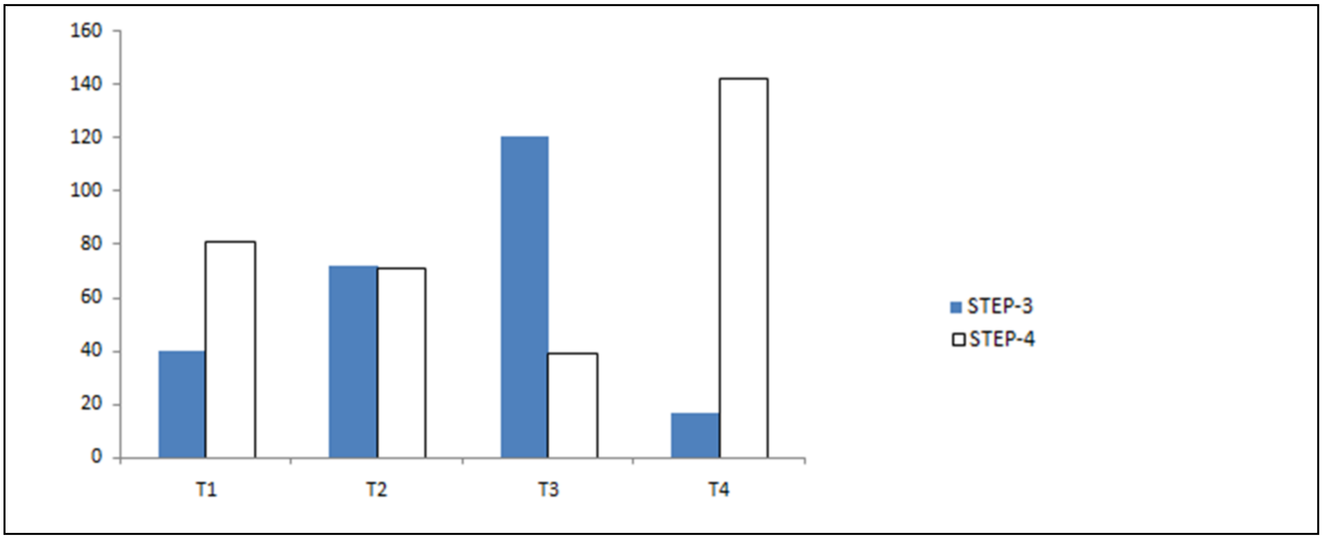


Fig. 4. Execution Time Required at different Conditions NB: - X-Axis Showing the different Task Assigned and Y-Axis Showing Completion Time Required by different Tasks.

VI. Conclusion

To realize the good performance of computing of scheduling of tasks in a cloud computing environment, a new algorithm is proposed. Different algorithms are tested for their suitability, feasibility, adaptability in the context of cloud scenario So that it can facilitate cloud-providers to provide a better quality of services. The proposed algorithm works on the problem exist when the number of small tasks is more in number or when the large tasks more in number. It performs in two phases. When the numbers of small tasks are more than the number of large size tasks then the algorithm will execute the large task first to increase efficiency to manage maximum completion time. In the reverse condition when the numbers of large size tasks are more than the number of small size tasks then the small tasks need to be executed first to increase the computing efficiency and to avoid starvation. In future the algorithm can be added with some other characteristics to enhance accountability.

References

[1] Tabak, E. K., Cambazoglu, B. B., & Aykanat, C. (2014). Improving the Performance of IndependentTask Assignment Heuristics MinMin, MaxMin and Sufferage. IEEE Transactions on Parallel and Distributed Systems, 25(5), 1244-1256. DOI:10.1109/tpds.2013.107.

[2] Nayak, B., Padhi, S. K., Pattnaik, P. K. (2017). Understanding the Mass Storage and Bringing Accountability. National Conference on Recent Trends in Soft Computing & It ´s Applications, pp. 28-35. ISSN: 2319 – 6734.

[3] Nayak, B., Padhi, S. K., & Pattnaik, P. K. (2018). Impact of Cloud Accountability on Clinical Architecture and Acceptance of Health Care System. 6th International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA). V.701, pp. 149-157. DOI 10.1007/978-981-10-7563-6_16.

[4] Suri, P. K. and Rani, R. (2017). Design of Task Scheduling Model for Cloud Applications in Multi-Cloud Environment. International Conference on Information, Communication and Computing Technology (ICICCT, Springer), 2017; 750:11–24.

[5] Mathew, T., Sekaran, K. C., & Jose, J. (2014). Study and analysis of various task scheduling algorithms in the cloud computing environment. 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). DOI:10.1109/icacci.2014.6968517.

[6] Brinkerhoff, D. W. (2004). Accountability and health systems: Toward conceptual clarity and policy relevance. Health Policy and Planning, 19(6), 371-379. DOI:10.1093/heapol/czh052.

[7] Singh, P., & Kaur, N. (2016). A Review: Cloud Computing using Various Task Scheduling Algorithms. International Journal of Computer Applications, 142(7), 30-32. DOI:10.5120/ijca2016909931.

[8] Banga, P., & Rana, S. (2017). Heuristic-based Independent Task Scheduling Techniques in Cloud Computing: A Review. International Journal of Computer Applications,166(1), 27-32. DOI:10.5120/ijca2017913901.

[9] Singh, S., & Kalra, M. (2014). Scheduling of Independent Tasks in Cloud Computing Using Modified Genetic Algorithm. 2014 International Conference on Computational Intelligence and Communication Networks. DOI:10.1109/cicn.2014.128.

[10] Reda, N. M. (2015). An Improved Sufferage Meta-Task Scheduling Algorithm in Grid Computing Systems. International Journal of Advanced Research, 2015; 3(10): 123 -129.

[11] Kumari, E., & Monika, A. (2015). Review On Task Scheduling Algorithms In Cloud Computing. International Journal of Science, Environment and Technology, 2015; 4(2): 433 – 439.

[12] Jain, N. S.,(2016). Task Scheduling In Cloud Computing using Genetic Algorithm. International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), 2016; 6(4): 9-22.

[13] Le, D., Bhateja, V., & Nguyen, G. N. (2017). A parallel max-min ant system algorithm for dynamic resource allocation to support QoS requirements. 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON). DOI:10.1109/upcon.2017.8251134.

[14] Nayak, B., Padhi, S. K., & Pattnaik, P. K. (2018). Static Task Scheduling Heuristic Approach in Cloud Computing Environment. 5th Springer International Conference on Information System Design and Intelligent Applications. Vol. 862, pp.473-480. DOI: 10.1007/978-981-13-3329-3_44.

[15] Rajput, S. S., & Kushwah, V. S. (2016). A Genetic Based Improved Load Balanced Min-Min Task Scheduling Algorithm for Load Balancing in Cloud Computing. 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN). DOI:10.1109/cicn.2016.139.

[16] Bao, L. N., Le, D., Nguyen, G. N., Bhateja, V., & Satapathy, S. C. (2017). Optimizing feature selection in video-based recognition using Max-Min Ant System for the online video contextual advertisement user-oriented system. Journal of Computational Science,21, 361-370. DOI:10.1016/j.jocs.2016.10.016.

Convolutional Neural Network Architecture for Plant Seedling Classification

Heba A. Elnemr
Computers and Systems Department
Electronics Research Institute
Cairo, Egypt

***Abstract*—Weed control is a challenging problem that may face crops productivity. Weeds are perceived as an important problem because they conduce to reduce crop yields due to the expanding competition for nutrients, water, and sunlight besides they serve as hosts for diseases and pests. Thus, it is crucial to identify weeds in early growth in order to avoid their side effects on crops growth. Previous conventional machine learning technologies exploited for discriminating crops and weeding species faced challenges of effectiveness and reliability of weed detection at preliminary stages of growth. This work proposes the application of deep learning technique for plant seedling classification. A new Convolutional Neural Networks (CNN) architecture is designed to classify plant seedlings at their early growth stages. The presented technique is appraised using plant seedlings dataset. Average accuracy, precision, recall, and F1-score are utilized as evaluation metrics. The results reveal the capability of the proposed technique in discriminating among 12 species (3 crops and 9 weeds). The system achieved 94.38% average classification accuracy. The proposed system is compared with existing plant seedling systems. The results demonstrate that the proposed method outperforms the existing methods.**

Keywords—Deep learning; convolutional neural network; plant seedling classification; weed control

I. Introduction

Plants remain an important and essential source of food and oxygen for nearly all living organisms on earth. Agriculture is prevailing in some continents like Africa, therefore appropriate automation of the farming procedure would assist in optimizing the crop yield and ensuring the perpetual productivity and sustainability. In accordance with [1], there is a sturdy bond between raised productivity and economic growth. Thus, the application of smart farming techniques in the agricultural sector can empower the development of the economy in many countries. Seedlings quality assessing proved to be a powerful means of prophesying the growth performance [2] and, hence, optimizing the plant production. Seedling classification is the first step to fulfill the seedling quality evaluation.

Furthermore, the invasion of weeds on farmlands leads to decline in the crop yield. Generally, weeds have no valuable beneficial, regarding nutrition, food or medication. However, they grow very quickly as well as they intrusively compete with original crops for space and nutrients [3]. Weeds identification is not an easy process due to the hazy boundaries of the crops, together with the diverse sandy and rocky backgrounds. Thus, there is a need to develop an efficient technique to accurately and certainly detect weeds from beneficial plants.

In order to improve agronomic production and crop quality, farmers should follow precision agriculture. Precision agriculture is a farm management approach that utilizes information technology and artificial intelligence to guarantee profit maximization, crop yield optimization, and environment preservation. One of the fundamental challenges that face precision agriculture is weed control. Weed control must be achieved earlier as possible after crop germination before weeds begin to compete with crops for nutrition and cause adverse effects. Thus, optimal weed treatment is recommended in the seedling stage. Nevertheless, in this phase, the discrimination between crops and weeds has some limitations; a) inadequate image resolution for distinguishing between exposed soil, crop seedlings and weeds, b) resemblance of spectra and appearances between weeds and useful crops in the early stages, and c) overlapping of the soil background reflectance with the detection process.

The application of machine learning techniques for automatic plant seedling classification has become a significant and promising field of research towards improving agriculture outcomes. Deep learning is a specific type of machine learning that has gained substantial interest in various disciplines. The Convolutional Neural Network (CNN) is a deep neural network architecture that is generally used to analyze visual images. Latterly, CNNs have achieved a significant breakthrough in computer vision fields. Additionally, the CNNs proved to have high ability to obtain the efficient features needed for image classification process [4]-[6].

Recently, CNNs have been broadly implemented in the agriculture domain for plant species identification [7]-[8], weed detection [9], and plant disease recognition [10].

In traditional image classification algorithms, handcrafted features are firstly extracted, then a feature selection process is achieved, and finally, a suitable classifier is chosen. However, CNN is proficient in learning various features from images, it covers global and local features, and it uses these features for efficient classification. CNN showed superior performance compared to other image processing techniques. Therefore, in this article, the enforcement of the CNN approach for plant seedling classification is investigated. The proposed system proceeds in four phases; preprocessing, constructing the network model architecture, training the network model and defining its parameters, and finally testing the designed network model. Despite the complexity of the acquired seedlings scenes due to illumination variations, resemblances between weeds and crops at premature stages of growth, and soil texture intricacies, the CNN succeeded to achieve high classification performance. Hence, this work aims specifically to develop a framework for crop-weed discrimination system that applies the CNN to classify 12 crops and weeds plant species and compare the proposed seedling classification system with other state-of-the-art techniques.

This article is structured as follows: Section 2 presents the related work. Section 3 details the CNN architecture devoted to developing the proposed deep plant seedling classification system. Section 4 describes and discusses the experimental results. Finally, Section 5 debates the conclusion and future work.

II. Related Work

Seedlings classification is a discipline that has got a substantial prominence in precision agriculture, since it permits for distant observation to the fields, providing a foundation for more efficacious weed control. Fine-grained weed control considerably depends on the accuracy of the classification process, so as the crops would not be damaged when treating the weeds. Accordingly, misclassification will possess a direct impact on crop yield.

In literature, classification of crop and weed species may be developed through two strategies. The first strategy is based on segmenting images into green and soil regions and extracts features from green patches and finally uses classification techniques to obtain the specified classes. The second strategy, on the other hand, relies on implementing deep learning techniques for plant seedling classification.

The work of [11] presented a method for classifying plant seedlings. This method aimed to improve the classification performance by consolidating the classification of the whole plants and the individual leaves. Thus, leaves are first separated from the plants then features are extracted from both the whole plants and the segmented leaves. The classification process is performed for the leaves and plants, and finally, Bayes belief integration is used to fuse the classification results. Bakhshipour and Jafari [12] applied two significant pattern recognition approaches; artificial neural networks (ANN) and support vector machine (SVM), to separate the weeds from the sugar beet plants using shape features. The shape features comprise Fourier descriptors and moment invariant features. Four species of prevalent weeds in the sugar beet fields were examined. The results indicate that SVM slightly outperforms the ANN. In [13] the authors developed a system vision technique relied on video processing as well as a hybrid ANN and ant colony algorithm classifier for assorting potato plant and three weed species. Texture features, obtained from the gray level co-occurrence matrix (GLCM) and the histogram, moment invariants, color features, and shape features are extracted. Then, the Gamma test is used to select the significant features.

Furthermore, spectral reflectance measurements are used for discriminating between crops and weeds [14] and [15]. In [14] an SVM along with spectral reflectance measurements are combined for developing a corn/silverbeet (as crop-weed) differentiation system. The intensities of the reflectance of laser beams off soil and vegetation at three wavelengths are gathered by a weed sensor. These reflectance measurements are used to compute the Normalized Difference Vegetation Indices (NDVIs). Two experiments are performed; in the first one, the obtained NDVI values are fed to an SVM to achieve the classification process, while in the second one, the raw reflected intensities are provided to the SVM for crop-weed discrimination. Strothmann et al. [15] proposed a crop-weed discrimination system based on in-field-labeling. A multi-wavelength laser line profile (MWLP) approach is used to scan plants and obtain spectral reflection intensities, scattering information at several wavelengths and 3D data. The spectral features are applied for separating soil and biomass, while the 3D surface features are exploited for discriminating crops and weeds.

The study of [16] investigates the classification of maize, weeds, and soil by training CNN to make a pixel-wise classification. The generated CNN is based on a modified architecture of the VGG16 at which the output layer is a convolutional layer instead of a fully connected layer. Eventually, semantic segmented images are obtained. Zhang et al. [9] proposed a system for identifying broad-leaf weeds in the pasture. Traditional machine learning techniques and deep learning approaches are investigated and compared. The results reveal that deep learning technique using CNN achieved high accuracy and robustness in detecting weeds in real-world pasture environments. The work [17] submitted an approach to classify the species of weeds and crops by employing CNN technique. The developed CNN is based on a hybrid network of AlexNet and VGGNET. The normalization notion is stimulated from AlexNet; while the filters' depth is selected based on VGGNET. Furthermore, incremental learning to learn new plant species is applied in this work.

III. Proposed CNN Architecture

In this work, CNN is adopted for plant seedling classification to automatically discriminate between weed species and crops at early growth stages. The proposed CNN consists of an input layer, hidden layers, and an output layer. The original seedling images are all equally resized to 128x128 pixels (this has been specified empirically such that to get satisfactory performance with acceptable processing speed) and fed to the input layer. The hidden layers consist of 5 stages of learning layers, as illustrated in Fig. 1. The utilized filters are all of kernel size 3x3 with a number of filters 32, 64, 128, 256 and 1024 for each convolutional layer within each stage, respectively.

The entire convolutional layers are associated with Rectified Linear Units (ReLU) layers, which apply the function f(x) = max (0, x) to the whole values of the input image. Thus, the negative input elements are set to 0. This decreases the training time and provides nonlinear rectifications, which escalates the nonlinear characteristics of the model and the whole network without impacting the receptive values of the convolutional layer [18].

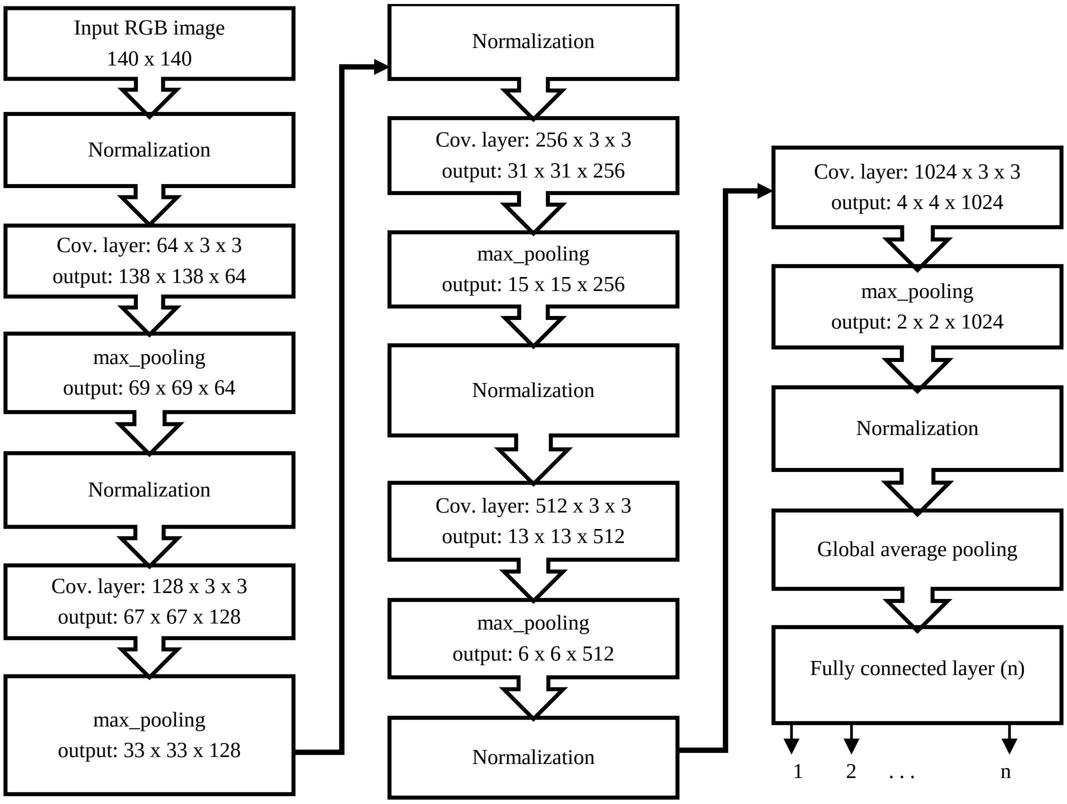


Fig. 1. The Proposed Deep CNN Architecture for Seedling Classification.

Each convolutional layer is followed by a pooling layer and a batch normalization layer. The pooling layer is utilized for reducing the output size of the layer before it, and hence, decreasing the computation complexity in the subsequent layers. A max-pooling procedure with a pool size of 2x2 is applied.

In deep CNN, small variations may augment as they pass through layers, which lead to change the distribution within each layer. This is called Internal Covariate Shift problem. Therefore, batch normalization is utilized to normalize each hidden layer inputs to stabilize their distribution and hence solve the Internal Covariate Shift problem. Furthermore, the batch normalization layer helps to faster the learning procedure [19].

Generally, the convolutional layers are used for feature extraction and the fully connected layers are used for classification tasks. Thus, the lower part of the CNN includes convolutional layers while the higher part comprises some fully connected layers. The fully connected layers have a large number of parameters which needs a high computational power and produces overfitting. On the other hand, the global average layer procedure computes the mean of each feature map and delivers it to the next layer. Hence, it does not need any parameter which minimizes overfitting [18]. Our proposed CNN architecture employs the global average pooling layer before the fully connected layers in order to reduce the utilized parameters and avoid overfitting.

In the output layer, the global average pooling layer is used to directly feed the obtained feature maps into the feature vectors. Finally, a fully connected layer, which comprises n (signifying the number of classes) nodes, along with softmax is realized to compute the probability of each predicted class.

The size of the output of each layer is declared in Fig. 1; after the normalization process, no change occurs in the output size.

IV. Experimental Results

A. Dataset

The utilized dataset, delivered by the signal processing group of the Aarhus University, in collaboration by Southern Denmark University, comprises 5539 images of roughly 960 unique plants categorized into 12 species (3 crops and 9 weeds) captured at early growth stages. It includes annotated RGB images with an approximate physical resolution of 10 pixels per mm.

Particularly, this dataset is adopted for researches that investigate plant species identification at their early germination stage. Thus, farmers (or robots for automatic weeding control) may be able to handle weeding before the weeds commence to compete with crops for nutrition. Additionally, the image segmentation process at this early stage is easier since the leaves have less overlapping in this stage [20]. The exploited dataset is detailed in Table I.

B. Evaluation Metrics

The proposed system performance is evaluated using the average accuracy, average precision, average recall, and average F1-score as follows.

TABLE. I. PLANT SEEDLINGS DATASET DETAILS

Class	Species	Training set images	Test set images	Total images
1	Black grass (Alopecurus myosuroides)	263	46	309
2	Charlock (Sinapis arvensis)	390	62	452
3	Cleavers (Galium aparine)	287	48	335
4	Chickweed (Stellaria media)	611	102	713
5	Common wheat (Tricicum aestivum)	221	32	253
6	Fat hen (Chenopodium 0album)	475	63	538
7	Loose silky-bent (Apera spica-venti)	648	114	762
8	Maize (Zea mays)	221	36	257
9	Scentless mayweed (Tripleurospermum perforatum)	516	91	607
10	Shepherd's purse (Capsella bursa-pastoris)	231	43	274
11	Small-flowered Cranesbill (Geranium pusillum)	490	86	576
12	Sugar beet (Beta vugaris)	385	78	463
Total		4738	801	5539

$$\text{average accuracy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\text{total number of correct samples for i}^{\text{th}}\text{ class}}{\text{number of total samples of i}^{\text{th}}\text{class}}\right) \quad (1)$$

$$\text{average precision} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\text{true positives}}{\text{true positives+false positives}}\right)_{\text{for i}^{\text{th}}} \quad (2)$$

$$\text{average recall} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\text{true positives}}{\text{true positives+false negatives}}\right)_{\text{for i}^{\text{th}}} \quad (3)$$

$$\text{average F1} - \text{score} = \frac{1}{n}\sum_{i=1}^{n} 2\times\left(\frac{\text{precision}\times\text{ recall}}{\text{precision+recall}}\right)_{\text{for i}^{\text{th}}} \quad (4)$$

C. System Evaluation

This section presents the evaluation strategy conducted through this work. The evaluation procedure is performed in two phases: training and testing. Thus, the seedling dataset is split into two separate sets; training and testing sets. In the training phase, 10-fold cross-validation is performed by randomly choosing 10% of the training set to represent the validation set, and this process is repeated for 10 successive rounds. In the training phase, the training images are used to fit the proposed CNN model and tune its hyperparameters while the validation images provide an unbiased appraisal of the model fitted on the training images during tuning its hyperparameters. On the other hand, in the test phase, the test images are utilized to afford an unbiased assessment of the final model that is fitted on the training dataset.

Extensive experimentations are achieved to evaluate the proposed method by comparing it with existing methods. These experiments are conducted by considering a different number of species. The first experiment involves 7 species (cleavers, chickweed, wheat, maize, scentless mayweed, Shepherd's purse and, sugar beet), the training set comprises 2472 images and the test set includes 430 images. The second experiment encompasses 8 species (charlock, cleavers, chickweed, fat hen, maize, scentless mayweed, Shepherd's purse, and sugar beet), the training phase contains (3116) images, whereas the testing phase holds (523) images.

The third experiment comprises 10 species (black grass, charlock, cleavers, chickweed, fat hen, loose silky-bent, maize, scentless mayweed, Shepherd's purse, and sugar beet), the training phase contains (4027) images, whereas the testing phase holds (683) images.

Finally, in the fourth experiment 12 species are considered and are divided into (4738) train images and (801) test images.

The proposed CNN is randomly initialized, and then it is trained for performing the classification process and indicated a convolutional model. The weights of the CNN are updated utilizing the training set. The final weights are selected using the validation set. For each iteration, the training and validation errors are computed, and the weights that achieve the minimum validation error are chosen.

Intel (R) Core (TM) i7-4702MQ CPU @ 2.20GHZ (8 GB RAM) processor is utilized for implementation. Python (Keras library) installed in Anaconda on the operating system Windows 10 is employed as a software tool for application.

D. Results

Table II presents the average validation performance of the proposed seedling classification for 7, 8, 10 and 12 species. It can be noticed from Table II that the validation accuracy reaches approximately 99% for all tested number of species, the validation recall, precision, and F1-score are roughly 98 % for 7 and 8 species while the validation recall reaches approximately 92% and 93% for the 10 and 12 species, respectively. In addition, the validation precision attains nearly 94% and 95% for 10 and 12 species, respectively, whereas the F1-score is about 93% for both 10 and 12 species.

On the other hand, Table III illustrates the average test performance. The results reveal that the average test accuracy, recall, precision, and F1-scale are approximately 99% for 7 species. For 8 species, the average test accuracy and recall are nearly 98% while the precision and F1-score are almost 99%. Furthermore, when using 10 species, the average test accuracy, recall, precision and F1-scale reach roughly 95%, 93%, 97 and 94, respectively. Finally, as testing the 12 species, the average test accuracy and F1-score has attained approximately 94%, while the average test recall and precision captured nearly 93% and 95%, respectively.

TABLE. II. THE AVERAGE VALIDATION PERFORMANCE OF THE PROPOSED SYSTEM FOR 7, 8, 10 AND 12 SPECIES

Number of plant species	**Accuracy (%)**	**Recall (%)**	**Precision (%)**	**F1-score (%)**	**Loss**
7 species	98.63	97.9	98.2	97.79	0.0701
8 species	98.91	97.56	97.68	97.58	0.051
10 species	98.76	92.24	94.02	92.73	0.0489
12 species	99.01	92.64	94.86	92.93	0.0565

TABLE. III. THE AVERAGE TESTING PERFORMANCE OF THE PROPOSED SYSTEM FOR 7, 8, 10 AND 12 SPECIES

Number of plant species	**Accuracy (%)**	**Recall (%)**	**Precision (%)**	**F1-score (%)**
7 species	98.61	98.63	98.92	98.77
8 species	98.28	98.38	98.59	98.47
10 species	94.88	93.22	96.47	94
12 species	94.38	93.1	94.83	93.57

Moreover, the confusion matrices of the proposed seedling classification method for 7, 8, 10, 12 species are displayed in Table IV, Table V, Table VI, and Table VII, respectively.

TABLE. IV. THE CONFUSION MATRIX FOR 7 SPECIES (1. CLEAVERS, 2. CHICKWEED, 3. WHEAT, 4. MAIZE, 5. SCENTLESS MAYWEED, 6. SHEPHERD'S PURSE AND 7. SUGAR BEET)

		Predicted classes						
		1	***2***	***3***	***4***	***5***	***6***	***7***
True classes	***1***	48	0	0	0	0	0	0
	2	0	100	0	0	2	0	0
	3	1	0	31	0	0	0	0
	4	0	0	0	36	0	0	0
	5	0	1	0	0	89	0	1
	6	0	0	0	0	1	42	0
	7	0	0	0	0	0	0	78

TABLE. V. THE CONFUSION MATRIX FOR THE 8 SPECIES (1. CHARLOCK, 2. CLEAVERS, 3. CHICKWEED, 4. FAT HEN, 5. MAIZE, 6. SCENTLESS MAYWEED, 7. SHEPHERD'S PURSE AND 8. SUGAR BEET)

		Predicted classes							
		1	***2***	***3***	***4***	***5***	***6***	***7***	***8***
True classes	***1***	60	1	0	0	0	0	1	0
	2	0	48	0	0	0	0	0	0
	3	0	0	101	0	0	1	0	0
	4	0	0	2	61	0	0	0	0
	5	0	0	0	0	36	0	0	0
	6	0	0	3	0	0	88	0	0
	7	0	0	0	0	0	1	42	0
	8	0	0	0	0	0	0	0	78

TABLE. VI. THE CONFUSION MATRIX FOR THE 10 SPECIES (1. BLACK GRASS, 2. CHARLOCK, 3. CLEAVERS, 4. CHICKWEED, 5. FAT HEN, 6. LOOSE SILKY-BENT, 7. MAIZE, 8. SCENTLESS MAYWEED, 9. SHEPHERD'S PURSE AND 10. SUGAR BEET)

		Predicted classes									
		1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
True classes	***1***	21	0	0	0	0	25	0	0	0	0
	2	0	62	0	0	0	0	0	0	0	0
	3	0	0	47	0	0	0	0	0	0	1
	4	0	0	0	100	0	0	0	2	0	0
	5	0	0	0	2	61	0	0	0	0	0
	6	2	0	0	0	0	112	0	0	0	0
	7	0	0	0	0	0	0	36	0	0	0
	8	0	0	0	2	0	0	0	89	0	0
	9	0	0	0	0	0	0	0	1	42	0
	10	0	0	0	0	0	0	0	0	0	78

TABLE. VII. THE CONFUSION MATRIX FOR THE 12 SPECIES (1. BLACK GRASS, 2. CHARLOCK, 3. CLEAVERS, 4. CHICKWEED, 5. WHEAT, 6. FAT HEN, 7. LOOSE SILKY-BENT, 8. MAIZE, 9. SCENTLESS MAYWEED, 10. SHEPHERD'S PURSE, 11. SMALL-FLOWERED CRANESBILL AND 12. SUGAR BEET)

		Predicted classes											
		1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
True classes	***1***	20	0	0	0	1	0	25	0	0	0	0	0
	2	0	60	0	0	0	0	0	0	0	0	2	0
	3	0	0	48	0	0	0	0	0	0	0	0	0
	4	0	0	0	100	0	0	0	0	2	0	0	0
	5	1	0	0	0	31	0	0	0	0	0	0	0
	6	0	0	0	2	1	60	0	0	0	0	0	0
	7	6	0	0	0	0	0	108	0	0	0	0	0
	8	0	0	0	0	0	0	0	36	0	0	0	0
	9	0	0	0	3	0	0	0	0	88	0	0	0
	10	0	0	0	1	0	0	0	0	1	41	0	0
	11	0	0	0	0	0	0	0	0	0	0	86	0
	12	0	0	0	0	0	0	0	0	0	0	0	78

E. Discussion

In this section, an inclusive debate for the realized outcomes and comparison with state-of-art are exhibited.

As considering 7 species (3 crops and 4 weeds) and 8 species (2 crops and 6 weeds) experiments, the evaluation results depict that the proposed technique has effectively and efficiently classified the experimented species. The suggested system achieved about 99% average accuracy, recall, precision, and F1-score of for 7species, and approximately 98% average accuracy and recall, as well as nearly 99% average precision and F1-score for 8 species.

On the other hand, for the 10 (2 crops and 8 weeds) and 12 (3 crops and 9 weeds) species, the performance evaluation is relatively less than that for 7 and 8 species. Despite the performance reduction, the classification evaluation still high and the empirical results manifest the potency and capability of the proposed system in discriminating among the various species. The submitted method, for 10 species, obtained average accuracy, recall, precision and F1-score of about 95%, 93%, 97%, and 94%, respectively. Additionally, for 12 species, the system attained approximately 94%, 93%, 95% and 94% average accuracy, recall, precision, and F1-score, respectively. It is apparent from Table VI and Table VII that Black grass and Loose silky-bent majorly affect the results due to their high similarity and insignificant differences at early-stage growth, and they are hard to be distinguished even by human eyes. As for other classes, the proposed model proved to be reliable and effective.

To scrutinize the performance of the proposed technique, a comparison is performed with some state-of-the-art. The proposed method is compared with the existing methods [21], [11] and [17].

The average accuracy of the existing seedling approaches and the proposed work are quoted in Table VIII. It may be observed that the proposed deep seedling classification system outperforms significantly [21] and [11]. Furthermore, it can be noticed that, in [11], the average accuracy for Cleavers and Fat hen classes are 81.4% and 81.6%, respectively, which are relatively weak. Yet, the proposed technique has developed it to 100% and 96.83%, respectively. Moreover, both works [11] and [21] skipped pretty similar species like black grass and loose silk bent. However, the submitted seedling classification approach achieves an average test accuracy of 94.38% for the whole 12 species.

TABLE. VIII. COMPARISON OF THE PROPOSED METHOD AND EXISTING METHODS

Method	**Number of species**	**Accuracy**
[21]	7	95.8
[17]	7	98.21
Proposed method	7	**98.61**
[11]	8	96.7
[17]	8	98.23
Proposed method	8	**98.28**
[17]	12	93.64
Proposed method	12	**94.38**

Over and above, the proposed system performance slightly exceeds that of [17] for 7 and 8 species, and significantly outperforms it for 12 species. Furthermore, it involves 6 convolutional layers and 3 fully-connected layers, whereas our proposed CNN comprises 5 convolutional layers and one fully connected layer. Adding more layers extends the number of hyperparameters, ergo the complexity of the system. Thus, the submitted CNN architecture is much simpler and provide superior performance.

V. CONCLUSION

In this article, a CNN architecture is developed to discriminate between plant images of crop species and weed species at several early growth stages. The proposed CNN has achieved an enhancement in performance owing to the combination of the presence of the normalization layer, the global average pooling layer and the choice of the depth of the filters. The results revealed that the elaborated CNN has an encouraging performance towards building a weed control system which is a step to precision agriculture. The proposed CNN model achieved average accuracy, recall, precision, and F1-score of 94.38, 93.1, 94.83, and 93.57, respectively, for discriminating 12 plant seedling (3 crops and 9 weeds). Furthermore, its architecture is simpler than other existing CNN models utilized for plant seedling classification. Additionally, its performance is much better than other existing methods.

In the proposed scheme, images that comprise single plant species are classified, thus, for classifying images with many plant species, the segmentation stage may be added to the system.

In addition, the proposed technique may be expanded to incorporate new plant species. Besides, the proposed technique may be implemented as a part of an IoT system for weed control, which can help to directly apply herbicides on the weeds without harming crops.

REFERENCES

[1] The Role of International Institutions in Economic Development and Poverty Reduction in the Developing World, Food and agriculture Organization of the United Nations. Rome, 2018. URLhttp://www.fao.org/3/I9900EN/i9900en.pdf

[2] L.O.L.A. Silva, M.L. Koga, C.E. Cugnasca, and A.H.R. Costa, “Comparative assessment of feature selection and classification techniques for visual inspection of pot plant seedlings”, Computers and Electronics in Agriculture vol. 97, pp. 47–55, 2013.

[3] D. Nkemelu, D. Omeiza, and N. Lubalo, “Deep convolutional neural network for plant seedlings classification”, Computer Vision and Pattern Recognition, arXiv: 1811.08404, 2018.

[4] A. Dhomne, R. Kumar, and V. Bhan,“Gender recognition through face using deep learning”, Procedia Computer Science, vol. 132, pp. 2-10, 2018.

[5] M. Radovic, O. Adarkwa, and Q. Wang, “Object recognition in aerial images using convolutional neural networks”, Journal of Imaging, vol. 3, no. 21, 2017.

[6] N. M. Zayed and H. A. Elnemr, “Intelligent Systems for Healthcare Management and Delivery”, chapter 5, “Deep Learning and Medical Imaging”, pp. 101-147, 2019.

[7] P. Bonnet, H. Goëau, S. Hang, M. Lasseck, M. Šulc, V. Malécot, P. Jauzein, J. Melet, C. You, and A. Joly, “Plant identification: experts vs. machines in the era of deep learning: deep learning techniques challenge flora experts”. In: A. Joly, S. Vrochidis, K. Karatzas, A. Karppinen, P. Bonnet (eds.). Multimedia Tools and Applications for Environmental &

Biodiversity Informatics. Multimedia Systems and Applications. Springer, Cham, Chapter 8, pp.131-149, 2018.

[8] Z. Qiu, J. Chen, Y. Zhao, S. Zhu, Y. He, and C. Zhang, “Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network”, Applied Sciences, vol. 8, no. 2, 212, 2018.

[9] W. Zhang, M. F. Hansen, T. N. Volonakis, M. Smith, L. Smith, J. Wilson, G. Ralston, L. Broadbent, and G. Wright, "Broad-Leaf weed detection in pasture," 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Chongqing, pp. 101-105, 2018.

[10] K. P. Ferentinos, “Deep learning models for plant disease detection and diagnosis”. Computers and Electronics in Agriculture, vol. 145, pp. 311–318, 2018.

[11] M. Dyrmann, P. Christiansen, and H. S. Midtiby, “Estimation of plant species by classifying plants and leaves in combination”, Journal of Field Robotics, vol. 35, pp. 202–212, June 2017.

[12] A. Bakhshipour, and A. Jafari, “Evaluation of support vector machine and artificial neural networks in weed detection using shape features”, Computers and Electronics in Agriculture, vol. 145, pp. 153–160, 2018.

[13] S. Sabzla, Y. A. Gilandeha, and H. Javadikia, “Developing a Machine Vision System to Detect Weeds from Potato Plant”, Journal of Agricultural Sciences, vol. 24, pp. 105-118, 2018.

[14] S. Akbarzadeh, A. Paap, S. Ahderom, B. Apopei, and K. Alameh, “Plant discrimination by Support Vector Machine classifier based on spectral reflectance”, Computers and Electronics in Agriculture, vol. 148, pp. 250–258, 2018.

[15] W. Strothmann, A. Ruckelshausen, J. Hertzberg, C. Scholz, and F. Langsenkamp, “Plant classification with In-Field-Labeling for crop/weed discrimination using spectral features and 3D surface features from a multi-wavelength laser line profile System”, Computers and Electronics in Agriculture, vol. 134, pp. 79–93, 2017.

[16] M. Dyrmann, A. K. Mortensen, H. S. Midtibya, and R. N. Jørgensen, “Pixel-wise classification of weeds and crop in images by using a Fully convolutional neural network”, International Conference on Agricultural Engineering 2016 - Aarhus University, Aarhus, Denmark, June 26–29, 2016.

[17] T. R. Chavan, and A. V. Nandedkar, “AgroAVNET for crops and weeds classification: A step forward in automatic Farming”, Computers and Electronics in Agriculture, vol. 154, pp. 361–372, 2018.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks”, Twenty-sixth Conference on Neural Information Processing Systems (NIPS), December 3-8, 2012.

[19] S. Ioffe, and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, The 32nd International Conference on Machine Learning (ICML 2015), Lille, France, July 6 –11, 2015.

[20] T. Giselsson, R. Jørgensen, P. Jensen, M. Dyrmann, and H. Midtiby, “A public image database for benchmark of plant seedling classification algorithms”, Computer Vision and Pattern Recognition, arXiv:1711.05458, November, 2017.

[21] P. Christiansen, and M. Dyrmann, “Automated classification of seedlings using computer vision”, Technical report, Aarhus University, Aarhus, 2014.

A Methodology for Engineering Domain Ontology using Entity Relationship Model

Muhammad Ahsan Raza[1], M. Rahmah[2], Sehrish Raza[3], A. Noraziah[4], Roslina Abd. Hamid[5]
Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Kuantan, Malaysia[1, 2, 4, 5]
Institute of Computer Science and Information Technology, The Women University, Multan, Pakistan[3]

***Abstract*—Ontology engineering is an important aspect of semantic web vision to attain the meaningful representation of data. Although various techniques exist for the creation of ontology, most of the methods involve the number of complex phases, scenario-dependent ontology development, and poor validation of ontology. This research work presents a lightweight approach to build domain ontology using Entity Relationship (ER) model. Firstly, a detailed analysis of intended domain is performed to develop the ER model. In the next phase, ER to ontology (EROnt) conversion rules are outlined, and finally the system prototype is developed to construct the ontology. The proposed approach investigates the domain of information technology curriculum for the successful interpretation of concepts, attributes, relationships of concepts and constraints among the concepts of the ontology. The experts' evaluation of accurate identification of ontology vocabulary shows that the method performed well on curriculum data with 95.75% average precision and 90.75% average recall.**

Keywords*—*Ontology engineering; semantic web; ontology validation; knowledge management

I. Introduction

Currently, the data is rapidly increasing and changing over the World Wide Web (WWW). In order to extract the precise information from a huge unstructured pool of WWW is like searching a needle in a haystack. For the extraction of precise and relevant information, researchers have proposed the representation of unstructured WWW data into an intelligent knowledge structure, namely, ontology. Ontology is a source of explicit specification of domain concepts, properties, constraints and security [1]. The ontology knowledge helps to uncover the implicit domain semantics which can be used in various intelligent systems such as query expansion [2, 3] and expert systems [4, 5].

In recent, various ontology building techniques have been developed and published by researchers and experts [6, 7]. These techniques provide useful guidelines for ontology creation such as ontology development life cycle, tools and ontology languages. Despite of the exiting techniques, the ontology development process is complex, restricted to particular scenarios, and lacks validation [9]. Moreover, domain ontologies are still inadequate and yet need to be applied widely [8]. To overcome these limitations, this research work proposed a simple and portable approach that facilitates ontology developers to create an accurate and quality based domain-specific ontology with fewer efforts and less time.

The proposed procedure of ontology engineering (OE) is novel in two aspects: (1) it is based on well know ER-schema which is readily available for most of the database-based organizations or can be developed efficiently for any domain of interest with accuracy. (2) The method proposes instant and cost-effective rules for ER to ontology translation while maintaining all semantic checks. The ER-schema and translation model (viewing the two facets as simple and portable) support the development of ontology for any knowledge domain. Focusing on the discipline of information technology where the learning material related to the curriculum is highly unstructured, we have developed an OE tool that captures semantics from the ER-schema and automatically constructs the domain ontology.

This research work has the following structure. Section 2 covers the background study which describes the exiting techniques related to OE. In Section 3 the proposed OE methodology is discussed. ER to ontology mapping model is outlined in Section 4. Section 5 gives the implementation detail of ontology development for the domain of information technology curriculum. Section 6 wraps up the article with the conclusion and future work.

II. Background Study

A. Ontology Engineering

OE deals with the systematic process of ontology creation. The technique describes the terms exploited in the domain and the associations between them. In the past, various manual OE techniques have been developed that are based on different steps [11]. However, many of the exiting methodologies confine the process of ontology creation into a group of different phases which are used only to create a native or domain-dependent ontology.

B. Issues in Ontology Engineering

In the process of ontology creation, it is necessary to discuss the three most important issues in detail.

1) Rules definition: The main reasons for ontology development include the removal of ambiguity among the ontology concepts, enlargement of ontology scope, and enhancement in the quality of domain knowledge. Furthermore, the procedure of ontology management turns into more complicated and multifaceted in large-scale development [12]. Therefore, to create a true and good quality ontology with smooth development procedure, and less effort and time, some rules (i.e., steps or phases) are needed to be defined and followed.

2) Reusability of domain information: An important goal of OE is to develop an ontology that can be used in various applications and tasks. Reusability not only saves time and effort but also increases the reliability and consistency of the ontology. The high reusability of ontology indicates its acceptance in various applications. For instance, a general ontology can be used to represent different domain aspects (e.g., UNSPSC[1]).

3) Ontology usability: The main focus of the ontology creation process is usability. Usability refers to an ontology that can accomplish the application requirements. Another point of view of usability is to create diverse or contradictory ontologies (satisfying application requirements) by using similar domain concepts. However, Gyrard et al. [13] argued that this feature makes the ontology dependent on particular application or task; thus, making its reusability low.

So, the technique involves in ontology creation must be based on good quality rules, and determine either ontology is useable or reusable. Note that if the ontology creation process pays attention to application requirements and utilization, the resulting ontology will be usable (application-dependent). On the contrary, the final ontology will become application-independent (reusable) if the development process overlooks the purpose and utilization of the application.

C. Literature Review

Currently, various techniques exist for the development of ontologies. Most of these techniques follow general steps: (1) identify the set of general terms, (2) create classes for the terms and then organize classes in a hierarchical formulation, and (3) finally apply constraints on identified associations between the classes. For instance, Reda et al. [14] have created an ontology graph (in RDF language) from the diverse Internet of Things (IoT) data to facilitate the interoperability of different IoT devices. The approach first constructs the IoT fitness ontology by recognizing the classes and associations between the classes using protégé ontology-editing tool. The proposed mapping rules and fitness ontology are then used to generate the semantic RDF graph (i.e., final ontology) for IoT. Another promising technique is adopted by [15], where the seven rules of software engineering and features of the structured design are combined to develop a generic educational ontology. However, the approach follows a complex procedure; comprising of six phases which are further split into smaller steps to build the final ontology. A similar approach is exploited in Remolona et al. [16], where machine learning and natural language processing techniques are combined to generate ontology from the journal database. Chujai et al. [20] have demonstrated a stepwise approach for building ontology from ER model using Protégé tool, whereas the research does not address the ontology translation independent of Protégé features. A more related approach is presented in [10], where a prototype tool is developed to automatically convert the relational data to ER schema. The intermediate ER data is then mapped to OWL ontology. However, the approach does not provide detail about the ontology development life cycle, and ER to OWL mapping (e.g., composite attribute and restriction mappings are inappropriate). The authors themselves suggest that the final ontology may be incomplete due to mapping inconsistencies. In recent, Ellefi et al. [17] proposed a novel model to develop ontology in culture heritage (CH) domain where the data is vast and diverse. The novelty is based on conceptualizing CH resources under three dimensions, namely, topology-based, photogrammetrical process-based and spatial information-based. Moreover, the authors have published the final ontology for knowledge sharing and reusability purposes.

Another vein of OE is to develop ontology from the textual data. In [18], authors have presented a novel technique for ontology construction. The method is based on the combination of knowledge extraction and knowledge capturing approaches from the text. The knowledge extraction approaches are used to extract the synonyms, terms, linear relations, hierarchical relationships and rules needed for ontology construction. On the other hand, latter methodologies (namely, natural language processing and text mining) provide a means to capture the semantic knowledge from textual data. Another model for extracting the vocabulary (terms and phrases) and semantic relations among the vocabulary from the agriculture textual data is presented in [19]. The model is based on RelExOnt algorithm for the automatic extraction of pre-defined relationships from the textual data. The final generated ontology is validated by experts against the limited relationships and achieved 75.7% precision.

In the literature survey, the main focus of this study was to analyze the core structure and ontology development life cycle of manual approaches towards OE. Furthermore, the general steps that every approach usually follow are also identified.

After the detailed analysis of existing methodologies, it has been observed that some techniques have not provided the detail of each step involved in ontology development life cycle, while others are specialized for a particular application only. Therefore, due to the absence of a standardized approach for OE researchers are still facing issues (as described in Section 2-B) in ontology development.

III. Engineering Ontology for Information Technology Curriculum

The proposed framework constructs ontology for Information Technology Curriculum (ITC) using the ER diagrams. The procedure for ontology development encompasses three simple phases: feasibility study, planning, and ontology formulation (as illustrated in Fig. 1). In the first phase, preliminary investigation is performed to gather the requirements for ontology domain. Based on the gathered data, ER-schema is crafted in the second phase. The last phase deals with the identification of ER schema to ontology (EROnt) conversion rules and implementation aspects of the system prototype.

A. Feasibility Study (Step 1)

For ontology construction, a preliminary investigation is performed to collect the entire requirements set of the ICT

[1] https://www.cs.vu.nl/~mcaklein/unspsc/

domain. This investigation helps to understand the domain and recognize the sources for acquiring knowledge about ICT.

1) Domain understanding: The ontology designers must have complete knowledge about the structure of the intended domain of interest. This knowledge is necessary to create an ontology in easy and smooth manner.

Ontology domain can be easily extracted by analyzing the detail of the targeted subject. For the ICT ontology, which must identify various courses to be taught under different disciplines of information technology, and relationships that exist between these disciplines, this research work has selected universities and Higher Education Commission (HEC) of Pakistan as our target subjects.

2) Gathering intended knowledge: To gather the requirements for ICT ontology, we have analyzed prospectuses and websites of numerous universities of Pakistan that follow the HEC curriculum. We have also consulted different experts and students from the universities to understand the hierarchy and structure of various disciplines, and courses taught in each discipline. Further, the latest edition of HEC curriculum is accessed that helped us to identify the classification of ICT courses (for instance, the learning material can be grouped into general-education, core, compulsory, supporting and elective categories).

B. Planning (Step 2)

Researchers utilize numerous methodologies to mine the knowledge for the ontology construction, for example, using a structured relational model [21] or exploiting an unstructured Web source [22] or utilizing both the structured and unstructured data sources [23]. The main objective of this step is to select the appropriate model to acquire the precise knowledge of the domain.

To achieve the goal of a simple and portable OE scheme to construct ontology, we have chosen the ER model. Benefits can be gained from using ER schema as (1) ER diagram is simple and can be created quickly with little expertise or by the experts of database systems, (2) many existing systems have already been archived in the form of ER diagrams (e.g., database management systems), and (3) ER is a portable model (not limited to a particular domain) that can accurately capture the conceptual needs of an intended domain (e.g., ICT). Fig. 2 depicts the example of ICT entities and relationships between them in which a class is denoted by a rectangle, relationship by diamond and attribute of the class by an oval.

C. Ontology Formulation (Step 3)

When the ICT entities, attributes, relationships, and cardinalities are identified in the form of ER-schema, the next step is to convert the schema into ontology knowledgebase. The process of ER to ontology translation can be viewed from Fig. 3.

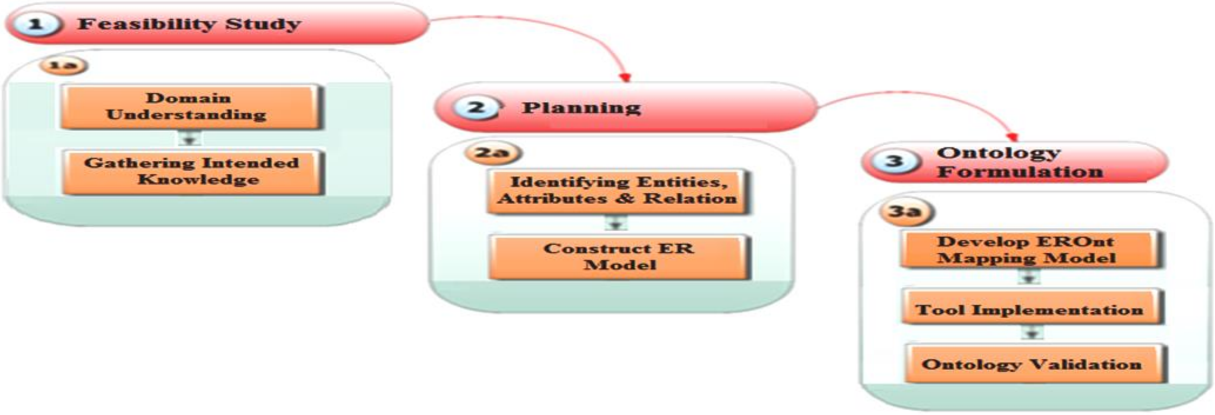


Fig. 1. Overview of Ontology Engineering Steps

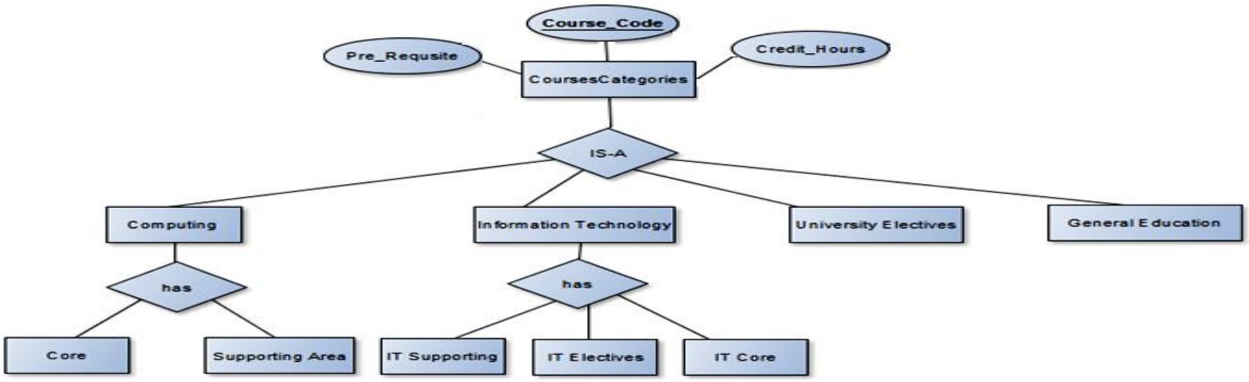


Fig. 2. Entity Relationship Diagram of Information Technology Curriculum.

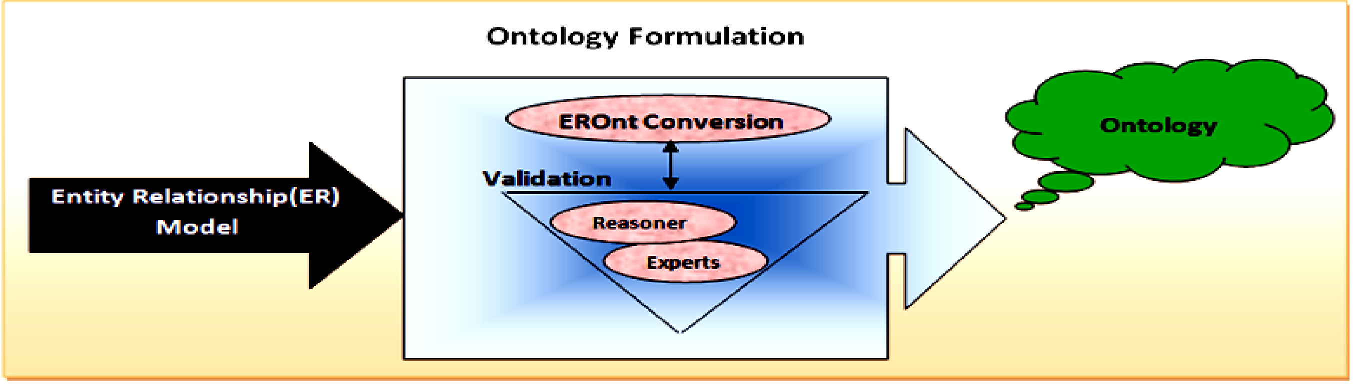


Fig. 3. The Procedure of Ontology Formulation.

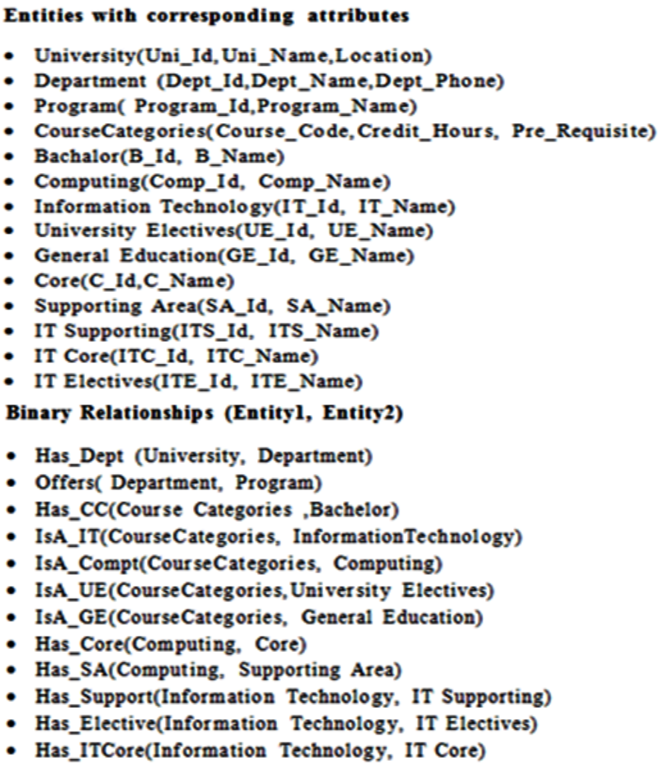
This phase proceeds with defining the EROnt conversion rules. These rules are then applied to illustrate the possible conceptualization (ontology) of ER model (see Section 4). Finally, the ICT ontology vocabulary (that includes concept data property, object property, and constraint) is validated using a reasoning engine (i..e., Herrmit). The use of semantic reasoner is a good choice as it enables to interpret logical consequences within the newly created ontology. This interpretation identifies the inaccurate and inconsistent classification of ICT concepts within the newly created ontology.

Further, the performance of EROnt translation model for the creation of ontology is evaluated in terms of precision and recall ratios (see Section 5). To this end, experts' opinion is obtained to assess whether the generated ontology vocabulary precisely reflects all the elements of ER model in ICT domain (i.e., developed in the planning phase of the proposed framework) or not.

IV. Interpreting EROnt Mapping Model

This section presents the ER model in detail and the proposed mapping rules for an ontology creation process. We have recognized a total of 54 entities, 12 relationships, and 31 attributes as a part of ICT ER schema. Fig. 4 gives a sample list of ICT entities along with their attributes, and relationships between the entities.

The semantic interpretation of ICT ER-schema associates each component of ER to ontology vocabulary (such as entity to a concept, attribute, and relationship to datatype property or object property, and cardinality to restriction) using OWL-Lite language [24]. Table I lists these interpretations as EROnt mapping rules for the ER to ontology conversion process. The conversion process proceeds by applying the set of outlined rules, for instance, entity or strong entity of ER schema is translated to OWL-class. Single-valued attribute having NULL value is mapped to OWL functional property while restricting the minimum cardinality to one. The final outcome of the process is the OWL ontology vocabulary that accurately mirrors all components of the ER schema.



Entities with corresponding attributes

- University(Uni_Id, Uni_Name, Location)
- Department (Dept_Id, Dept_Name, Dept_Phone)
- Program(Program_Id, Program_Name)
- CourseCategories(Course_Code, Credit_Hours, Pre_Requisite)
- Bachalor(B_Id, B_Name)
- Computing(Comp_Id, Comp_Name)
- Information Technology(IT_Id, IT_Name)
- University Electives(UE_Id, UE_Name)
- General Education(GE_Id, GE_Name)
- Core(C_Id, C_Name)
- Supporting Area(SA_Id, SA_Name)
- IT Supporting(ITS_Id, ITS_Name)
- IT Core(ITC_Id, ITC_Name)
- IT Electives(ITE_Id, ITE_Name)

Binary Relationships (Entity1, Entity2)

- Has_Dept (University, Department)
- Offers(Department, Program)
- Has_CC(Course Categories ,Bachelor)
- IsA_IT(CourseCategories, InformationTechnology)
- IsA_Compt(CourseCategories, Computing)
- IsA_UE(CourseCategories, University Electives)
- IsA_GE(CourseCategories, General Education)
- Has_Core(Computing, Core)
- Has_SA(Computing, Supporting Area)
- Has_Support(Information Technology, IT Supporting)
- Has_Elective(Information Technology, IT Electives)
- Has_ITCore(Information Technology, IT Core)

Fig. 4. Schema of Information Technology Curriculum.

TABLE. I. EROnt Mapping Model

ER Components			Ontology Components
Entities	Entity / Strong entity		- Map to class
	Weak entity		- Transform as a subclass of class obtain from a strong entity called host class - Set the subclass min-cardinality to one and object-property in the host class with the range set to subclass class
Attributes	Attribute		- Map to data-type property
	Key Attributes		- Map to functional data-type properties - Max-cardinality is set to one and the uniqueness is represented by an inverse functional property
	Data Type (date, varchar, integer etc.)		- Map to ranges of data-type property
	Single value attribute	Null	- Map to a functional data-type property with min-cardinality set to zero
		Non-Null	- Map to a functional data-type property with min-cardinality set to one
	Composite Attribute		- Ignore the composite attributes and map simple attributes into data-type properties (OR) Map composite attributes as a sub-properties of corresponding data-type properties
	Multi-valued Attribute	Null	- Map to a data-type property with min- cardinality set to zero
		Non-Null	- Map to a data-type property with min- cardinality set to one
Relations	Relation		- Map to an object-property
	IS-A relationship		- Map to subClassOf relation
	Ternary relationship		- Map relation as a class having three inverse object-properties (participating class, associating class and relationship class)
	Recursive Relation (constraints)	1:N relationship	- Map to a object-property with range and domain set to the same class - Set min and max cardinalities
		M:N relationship	- Map relation to a class: an entity and relationship class with someValueFrom constraint
	Binary Relation (with Attributes)		- Map relation to a class and create two object-properties (relation class and participating class)
	Binary Relation (No Attributes)		- Map to two object-properties which are inverse of each other
	Binary Relation (constraints)	1:M relationship	- Map to min and max cardinalities
		M:N relationship	- Apply constraints after dividing relationship into 1: M and M: 1 relationships
		1:1 relationship	- Map as a functional property and set max- cardinality to one

V. Implementation and Results

The ICT-schema and proposed EROnt rules are used to obtain the ontology vocabulary for an intended domain. A system prototype is developed using the Java framework and language. Other tools and APIs such as ARQ engine, and Jena API are used to implement the ER to ontology translation to obtain the resultant ICT OWL-ontology. The proposed method also utilized Protégé (an open-source ontology editor) to visualize and validate the resultant ontology. Fig. 5 from the Ontograf tool (a visual built-in plug-in in protégé) depicts a graphical overview of the new ICT ontology. In order to validate the consistency of ICT ontology, this study relied on logic reasoning engine, namely, Hermit (i.e., plug-in in protégé). The reasoner tested the ontology (without human intervention) for concepts redundancy and accuracy of extracted relationships between the concepts, and reported consistency of 100%.

Furthermore, the identified vocabulary (e.g., concepts, relationships) of new ontology is inspected by the experts to estimate the performance of the system prototype. Two groups of twenty participants (i.e., a faculty member and research students) from two universities have taken part in the evaluation process. Each group of experts received an ER schema and the corresponding generated ontology to explore four key ontology elements: (1) concepts, (2) data property, (3) object property and (4) constraints, that was obtained as an outcome of the system prototype. Furthermore, the group assessment was shuffled with each other to avoid any miss-interpretation.

We have calculated precision and recall values to measure the effectiveness of the system prototype. The precision measure is important as it represents the accurate modeling of domain knowledge, while recall value shows the system reliability in EROnt rules to generate the final ontology. These measures are calculated manually from the experts' judgment about the extracted ontology vocabulary using Equation (1) and Equation (2) as follows:

$$Precision = \frac{valid\ \ number\ of\ T\ extracted}{Total\ number\ of\ T\ extracted} \quad (1)$$

$$Recall = \frac{valid\ number\ of\ T\ extracted}{Total\ number\ of\ T\ present\ in\ ER\ model} \quad (2)$$

where *T* can be either concept, attribute or relation.

Table II reports the results for the extracted vocabulary. From the results, it is evident that our approach achieved high value for precision measure (i.e., valid vocabulary identification). Recall findings are also significant with a little variation in the reliable conversion of constraints, which we believe is might be because of inconsistency in the design of ICT ER-schema. Ultimately, the framework achieved 95.75% average precision and 90.75% average recall in the overall procedure of engineering the ICT domain ontology.

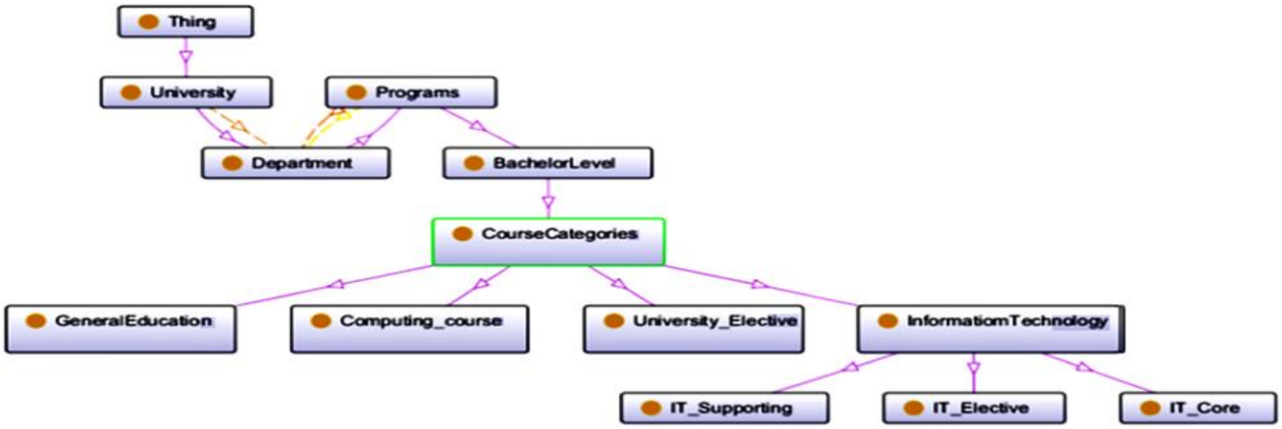


Fig. 5. Snapshot of Resultant Ontology.

TABLE. II. PRECISION AND RECALL OF EXTRACTED VOCABULARY

Vocabulary	Evaluation Measures	
	Precision	Recall
Concepts	0.98	0.94
Data properties	0.96	0.92
Object properties	0.96	0.91
Constraints	0.93	0.86
Average result	0.9575	0.9075

VI. CONCLUSION AND FUTURE DIRECTIONS

The use of ICT ontology can improve domain description and meaningful information retrieval. This semantic structure facilitates students in course selection as well as researchers in identifying the constitution and hierarchies of higher education curriculum. However, the formulation of an ontology requires the acquisition of complete and precise description of the ICT structure. Furthermore, it is important that OE process is done efficiently and accurately.

Keeping in view, we have presented the ER-schema based approach that allows researchers to develop a domain ontology in standard and domain-independent form. In the context of ICT, our methodology acquires ICT needs from the universities and HEC documentation. The ER model of ICT is used as a representation of domain requirements due to its semantic orientation. The ontology vocabulary (concepts, properties, etc.) is then identified from the ER schema using EROnt translation rules. These rules influence the working of system prototype in the overall process of OE. The evaluation via experts (in terms of precision and recall) and a logic reasoner (i.e., consistency test) confirm that the resultant ICT ontology accurately represents the domain knowledge.

In the future, the ICT ontology can be enhanced by adding other disciplines and constraints which make its use feasible for every field of academia, and for the users in semantic search over WWW.

ACKNOWLEDGMENT

The authors would like to express the gratitude to University Malaysia Pahang for funded this research under RDU 190112.

REFERENCES

[1] G. K. Saha, “Web Ontology Language (OWL) and semantic web,” ACM Ubiquity, vol. 8, pp. 1-24, September 2007.

[2] M. A. Raza, R. Mokhtar, N. Ahmad, and A. Mahmood, “Sensual Semantic Analysis for Effective Query Expansion,” International Journal of Advanced Computer Science and Applications (IJACSA), 9(12), 2018. http: //dx.doi.org/10.14569 /IJACSA.2018.091208.

[3] M. A. Raza, R. Mokhtar, N. Ahmad, R. A. Hamid , F. Zainuddin and N. A. Ahmad, “Query Expansion Using Conceptual Knowledge in Computer Science,” Advanced Science Letters, 24 (10). pp. 7490-7493, 2018, ISSN 1936-6612.

[4] R. A. Hamid, N. A. Ahmad, R. Mokhtar, F. Zainuddin and H. L. K. Huat, "Herbal Identification and its Benefits using Production Rules Approach," In: Proceeding of International Competition and Exhibition on Computing Innovation 2016, 6-8 December, Universiti Malaysia Pahang, pp. 206-211, 2016, ISBN 928-967-2054-04-7.

[5] R. A. Hamid, R. Mokhtar, N. A. Ahmad, F. Zainuddin and A. Abdullah, "i-Herbs: An Expert System for Malaysian Herbs Identification Using Production Rules Approach," Advanced Science Letters, 24 (10). pp. 7815-7818, 2018, ISSN 1936-6612.

[6] R. Nilsson Hall and A. Jerjas, ‘Specifying an ontology framework to model processes in hospitals’, Dissertation, 2017.

[7] I. Harrow, “Ontologies Guidelines for Best Practice,” Pistoia Alliance Public resources, December, 2016.

[8] M. A. Raza, R. Mokhtar, N. Ahmad, M. Pasha, and U. Pasha, “A Taxonomy and Survey of Semantic Approaches for Query Expansion,” IEEE Access, 17823-33, 7, 2019.

[9] M. Hepp, “Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies,” IEEE Internet Computing, 11, 90-96, 2007.

[10] C.-h. Liao, Y.-f. Wu and Guo-hua King, "Research on Learning Owl Ontology from Relational Database," Journal of Physics: Conference Series, 1176, 022-031, 2019.

[11] U. Yadav, G. S. Narula, N. Duhan, and V. Jain, “Ontology Engineering and Development Aspects: A Survey,” I.J. Education and Management Engineering, 3, 9-19, 2016. doi: 10.5815/ijeme.2016.03.02.

[12] A. M. Khattak, K. Latif, S. Lee, and Y.-K. Lee, “Ontology Evolution: A Survey and Future Challenges', in U- and E-Service, Science and Technology,” ed. by Dominik Ślęzak, Tai-hoon Kim, Jianhua Ma, Wai-Chi Fang, Frode Eika Sandnes, Byeong-Ho Kang and Bongen Gu (Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 68-75, 2009.

[13] A. Gyrard, M. Serrano, and G. A. Atemezing, "Semantic Web Methodologies, Best Practices and Ontology Engineering Applied to Internet of Things," in 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), pp. 412-17, 2015.

[14] R. Reda, F. Piccinini, and A. Carbonaro, “Towards Consistent Data Representation in the Iot Healthcare Landscape,” in Proceedings of the 2018 International Conference on Digital Health, Lyon, France, pp. 5-10, 2018.

[15] Z. Luo, M. Deng, and L. Yongjian. "An Ontology Construction Method for Educational Domain," Intelligent Systems Design and Engineering Applications, 2013 Fourth International Conference on. IEEE, 2013.

[16] M. F. M. Remolona et al., “Hybrid Ontology-Learning Materials Engineering System for Pharmaceutical Products: Multi-Label Entity Recognition and Concept Detection,” Computers & Chemical Engineering, 107, 49-60, 2017.

[17] M. B. Ellefi et al., “Cultural Heritage Resources Profiling: Ontology-Based Approach,” in Companion Proceedings of the The Web Conference (Lyon, France: International World Wide Web Conferences Steering Committee, 2018), pp. 1489-96, 2018.

[18] P. Buitelaar, P. Cimiano, B. Magnini, “Ontology learning from text: an overview, Ontology Learning from Text: Methods, Evaluation and Applications,” IOS Press, Amsterdam, The Netherlands, pp. 1–10, 2005.

[19] N. Kaushik, and N. Chatterjee, “Automatic Relationship Extraction from Agricultural Text for Ontology Construction,” Information Processing in Agriculture, 5, 60-73, 2018.

[20] P. Chujai, N. Kerdprasop, K.Kerdprasop, “On Transforming the ER Model to Ontology Using Protégé OWL Tool,” International Journal of Computer Theory and Engineering, Vol. 6, pp. 484-489, 2014.

[21] M. Li, X. Du, and S. Wang, "A Semi-Automatic Ontology Acquisition Method for the Semantic Web," Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 209-20, 2005.

[22] V. Subramaniyaswamy, "Automatic Topic Ontology Construction Using Semantic Relations from Wordnet and Wikipedia," Int. J. Intell. Inf. Technol., 61-89, 2013.

[23] F. Getahun, and Kidane Woldemariyam, “Integrated Ontology Learner: Towards Generic Semantic Annotation Framework,” in Proceedings of the 9th International Conference on Management of Digital EcoSystems (Bangkok, Thailand: ACM), pp. 142-49, 2017.

[24] G. Antoniou and F. V. Harmelen, “Web ontology language: Owl,” Handbook on Ontologies, Springer Berlin Heidelberg, pp. 67-92, 2004.

An Evaluation Model for Auto-generated Cognitive Scripts

Ahmed M. ELMougi[1], Yasser M. K. Omar[3]
College of Computing and Information Technology
Arab Academy for Science, Technology and Maritime Transport, Cairo, Egypt

Rania Hodhod[2]
TSYS School of Computer Science
Columbus State University
GA, USA

***Abstract*—Autonomous intelligent agents have become a very important research area in Artificial Intelligence (AI). Socio-cultural situations are one challenging area in which autonomous intelligent agents can acquire new knowledge or modify existing one. Socio-cultural situations can be best represented in the form of cognitive scripts that can allow different techniques to be used to facilitate knowledge transfer between scripts. Conceptual blending has proven successful in enhancing the social dynamics of cognitive scripts, where information is transferred from similar contextual scripts to a target script resulting in a new blended script. To the extent of our knowledge, there is no computational model available to evaluate these newly generated cognitive scripts. This work aims to develop a computational model to evaluate cognitive scripts resulting from blending two or more linear cognitive scripts. The evaluation process involves: 1) using the GloVe similarity to check if the transferred events conceptually fit the target script; 2) using the semantic view of text coherence to decide on the optimal position(s) to place the transferred event(s) in the target script. Results show that the GloVe similarity can be applied successfully to preserve the contextual meaning of cognitive scripts. Additional results show that GloVe embedding gives higher accuracy over Universal Sentence Encoder (USE) and Smooth Inverse Frequency (SIF) embedding but this comes with a high computational cost. Future work will look into reducing the computational cost and enhancing the accuracy.**

***Keywords*—Autonomous intelligent agents; socio-cultural situations; cognitive scripts; conceptual blending; contextual structural retrieval algorithms; text coherence; sentence embedding**

I. Introduction

Autonomous intelligent agents are a very important research area in Artificial Intelligence (AI). Intelligent agents possessing mental abilities, such as knowledge, belief, intention, and obligation can have human-like capabilities, such as artificial intuition and imagination, analogy and conceptual blending, design, writing poetry, argumentation, dialogue generation, negotiation abilities and shared mental models. It is important for autonomous intelligent agents to be able to acquire new knowledge or modify existing one. This seems to be quite difficult in some domains, such as socio-cultural situations because of the temporal and causal relations twined in these situations. Generally, people think of a situation as a sequence of routine actions/events that can be represented in the form of cognitive scripts. These events are connected temporally or causally with preceding and succeeding events [1].

It is a challenge to develop an intelligent agent that has the mechanism to change the knowledge it has and learn from external knowledge. Humans use analogical reasoning [2] to learn by simply transferring knowledge from a more familiar situation to a less familiar one making use of the structural similarity of the two situations. Conceptual blending is a theory of cognition, developed by Gilles Fauconnier and Mark Turner [3], that uses analogical reasoning to enhance the social dynamics of one script (target) by transferring events from a contextually similar script (base) resulting in a new blended script [1], [4], [5], [6].

One shortcoming that can be seen in these works is the fact that the evaluation of the resulting scripts needs human intervention; there should be a computational model to evaluate the newly generated blended scripts particularly in real time applications such as interactive narrative applications [4], [5]. Some events may be transferred to the target script while they don’t conceptually fit it. For example, the event “audience listens to movie” may be transferred to the cinema script when blending it with the lecture script. An event may be inserted into an unlogic position in the target script. For example, the event “light on” may be inserted before the event “movie starts” in the cinema script, when blending it with the lecture script, while it must be inserted after the event “movie ends”.

This work aims to develop a computational evaluation model for blended linear cognitive scripts. The approach used in [1] and [6] is adopted in this work to select events to be transferred from the base script to the target script. The evaluation process is two phases: Firstly, checking if the selected events can be added to the target script; The GloVe similarity ratio between every selected event and the target script is computed. If this ratio exceeds or equals a specified threshold, the event can be added to the target script. Secondly, using text coherence evaluation techniques is to specify the optimal position(s) to insert the selected event(s). The target script is converted into a text with every event converted into a sentence. The transferred events are converted into sentences. The optimal positions of the transferred events are the ones with the highest semantic text coherence [7]; in this technique, every sentence is converted into a vector and the semantic similarities between every two subsequent sentences are computed and then averaged to compute the coherence of the text. The semantic similarity between two sentences is the cosine similarity between their corresponding embedding vectors.

This paper is organized as follows: Section II gives background about cognitive scripts and analogical reasoning. Section III presents related work in the field of enhancing the dynamics of socio-cultural situations highlighting approaches in the fields of text coherence evaluation techniques and sentence embedding. Section IV introduces the proposed model. Section V shows results discussion. Section VI provides conclusion. Section VII provides suggestions for future work.

II. Background

A. *Cognitive Scripts*

We live in a world consisting of objects and events that relate these objects to each other. People store their knowledge about socio-cultural situations as a sequence of events, such as “Entering a restaurant” or “Attending a lecture” situations. Such socio-cultural situations are best represented in the form of cognitive scripts with events connected by directional edges. The events are either temporally or causally connected in a way that defines the context of a cognitive script. A cognitive script may be linear or multi-branched as shown in Fig. 1 in which each path in the multi-branched script can be seen as an independent linear script [1]. In this figure, the cinema cognitive script consits of different events such as “Audience buys ticket”, “Lights off”, and “Audience watches movie”. These events are conneted to their preceding and succeeding events by directional edges. The script consists of four different paths, each path represent a linear script. These paths have three interscting events; “Audience enters auditorium”, “audience watches movie” and “Movie ends”.

B. *Analogical Reasoning*

Analogical reasoning is a core process in human cognition defined as the ability to perceive and use relational similarity between two situations. In analogical reasoning, the relational similarity between two situations can be used to make inferences from one situation to the other. The first situation is called the base situation and is more familiar than the second situation which is called the target situation [2].

Analogical reasoning can be applied to production rules, cases, semantic networks, and cognitive scripts [8] and is usually comprised of three stages; retrieval, mapping, and evaluation. Retrieval is the process of retrieving a situation from long-term memory that is analogous to a situation in working memory. Mapping is the core process in analogical reasoning and is defined as the process of finding structural similarity between two situations and making inferences from a base situation to a target situation. Two situations can be structurally similar if there is an alignment between the two situations according to their structural similarity. Only then projected inferences from a base situation to a target situation can occur noting that every object in the base situation must be aligned to only one object in the target situation. This is known as one-to-one-correspondence. Sterman and his colleagues at MIT made an interesting analogy between the inflow and outflow of water in a bathtub with CO_2 emissions and removal in the atmosphere. In this analogy, the bathtub corresponds to the atmosphere. Water inflow and water outflow correspond to CO_2 emissions into the atmosphere and CO_2 removal respectively. Another requirement for structural consistency of two situations is that when two relations are matched, their arguments must be matched. Finally, inference from the base situation to the target situation is selective. People prefer to infer relations that are consistent with the matching structure of the two situations, in addition to using the systematicity principle. Lastly, evaluation takes place where analogy and inferences are accepted or rejected. Three factors affect the evaluation: The first factor is factual correctness that clarifies whether inferences are true or not. This may be incorrect in the case of future predictions. Another aspect related to factual correctness is adaptability which means that inferences can be accepted if they can be adapted easily in the target situation. The second factor is goal relevance and is important in problem-solving situations. The third factor is related to whether new knowledge can be added to the target situation or not. This may be risky, but it is important in brainstorming or unfamiliar situations [2].

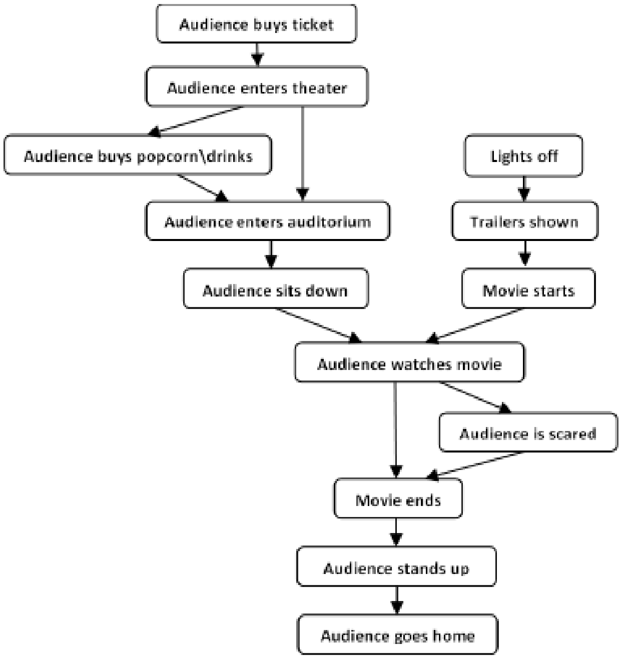


Fig. 1. Multi-Branched Cinema Cognitive Script.

III. Literature Review

A. *Related Work*

Many works have been done to create new knowledge from existing one in socio-cultural situations.

Hodhod and Magerko (2014) tried to give AI improvisational agents the ability to improvise new non-traditional scenes making use of existing social cognitive scripts [4]. Improvisational acting is a creative process implemented by actors on stage in real time. In this process, actors use their perceptions of the environment to create stories with each other. The authors developed the Pharaoh algorithm that retrieves a contextually similar cognitive script (base) to be blended with a target script based on the events’ appearances in the scripts and their least common parents [9]. iPharaoh, a modification of the Pharaoh algorithm, was after then developed to enhance the performance of the Pharaoh algorithm in terms of precision and recall, in addition to reducing the retrieval time [6]. In this work, a target script is chosen from the script-base. The Pharaoh algorithm is then used to find the highest contextually similar script among the script-base, referred to as the base script. The same conceptual blending rules used in the cognitive system, Sapper [10], are applied to the target and base scripts. These rules keep the structure of the target script and allow the addition of new events from base script or semantic networks.

The main drawbacks in Pharaoh and iPharaoh are that both algorithms rely on exact matching, in addition to the absence of a computational evaluation model; the blended scripts are evaluated by humans.

Permar and Magerko (2013) used another approach in [5]. The authors were interested in scripts in the domain of pretend play. The scripts are represented using a Directed Acyclic Graph (DAG). The authors used a blending algorithm that consists of three phases; counterpart mapping, mapping selection, and mapping application. Node mapping in the counterpart phase is a key process, the priority for node mapping followed the following rules:

- If the path of the target script has an iconic node, this node must be replaced.
- If the path of the base script has an iconic node, it has a priority to replace the mapped node in the target script.
- All pairs of paths are ordered according to the number of mapped nodes from the highest to the lowest.
- If a node is selected in a path, it will not be selected in the succeeding paths.

Some of the drawbacks in this approach are the use of exact matching, blending can take place between two scripts only, causality is not considered, and blended scripts are evaluated by humans.

Gawish et al. (2013) modified Pharaoh to allow the use of WordNet for lexical similarity [1]. WordNet is a lexical database with nouns and verbs organized into hierarchies of an *is–a* relation [11]. This representation makes WordNet particularly suited for similarity measures between two distinct but similar words. The model used consists of four blocks; evolved script-base, retrieval module, commonsense knowledge base, and learning module. The retrieval module retrieves the base script, which is the one that has the highest contextual similarity with the target script. The learning module uses two evolutionary processes to create a blended script; crossover and mutation. Crossover is used to insert new events or connections from the base script into the target script. Mutation is used to insert new events or connections from ConceptNet. ConceptNet is a large-scale commonsense semantic network of assertions of commonsense knowledge that represents the spatial, physical, social, temporal, and psychological aspects of everyday life [12]. Although this work addressed some of the drawbacks in the previous works, such as the use of lexical similarity and the ability to learn from other scripts as well as commonsense knowledge, it still did not address the missing ability of automatic evaluation of the resulting blended scripts.

B. Text Coherence Evaluation Techniques

Thinking of a cognitive script as a sequence of text can be the starting point to allow the emergence of an automated evaluation model. Testing coherence is to specify if the text is well written or not can be used to evaluate cognitive scripts. Text coherence evaluation has been used in many applications, such as machine translation, text generation, and summarization. Two important approaches in text coherence evaluation are introduced in [7]. The first approach is based on the syntactic view of text coherence which considers the change of the syntactic role of the text entities through adjacent sentences based on the Centering Theory. This theory asserts that texts in which successive statements mention the same entities are more coherent than texts in which multiple entities are mentioned.

The other approach relies on the semantic view of text coherence which implies that coherent text has high lexical cohesion between its sentences. This means that subsequent sentences have high semantic similarity. The semantic similarities between every two subsequent sentences are measured and then averaged to get the text coherence. This is illustrated in (1).

$$coherence(T) = \frac{\sum_{i=1}^{n-1} sim(S_i,S_{i+1})}{n-1} \tag{1}$$

Where *sim(S*$_i$*,S*$_{i+1}$*)* is the measure of semantic similarity between sentences *Si* and *Si*$_{+1}$.

The authors experimented with three different approaches to measure the semantic similarity between two sentences. The first approach measures the semantic similarity of two sentences in terms of word overlap. This is illustrated in (2).

$$sim(S_1, S_2) = \frac{2|words(S_1) \cap words(S_2)|}{(|words(S_1)| + |words(S_2)|)} \tag{2}$$

Where *words(S*$_i$*)* is the set of words in sentence *i*. The main drawback of this approach is that two sentences may have no common words, but they are semantically related. For example, the sentences “game ends” and “audience stands up” have no common words although they are semantically related.

The second approach is to use WordNet similarities between words in the two sentences. Since the WordNet similarity between words is dependent on the meaning of the words, the higher the similarity value between two words is, the more similar these words are. This is illustrated in (3):

$$sim(S_1, S_2) = \frac{\sum_{\substack{w_1 \in S_1 \\ w_2 \in S_2}} \underset{\substack{c_1 \in senses(w_1) \\ c_2 \in senses(w_2)}}{argmax} sim(c_1,c_2)}{|S_1||S_2|} \tag{3}$$

Where |S$_i$| is the number of words in sentence *i*. Since the appropriate senses of words w_1 and w_2 are not known, the similarity measure will select the senses which will maximize *sim(c*$_1$*,c*$_2$*)*.

One concern in this approach is the possibility of the words with high WordNet similarity to be irrelevant to the context/meaning of the text in the cognitive scripts.

The third approach follows the method in [13] and converts every sentence into an embedding vector then it measures the cosine similarity between the two embedding vectors. This approach is the approach used in the evaluation model used in this work.

C. Sentence Embedding

Different word and sentence embeddings have been developed to encode words and sentences as numerical vectors to be used in different natural language processing applications. Examples of these embeddings which will be used in this work are GloVe, Universal Sentence Encoder (USE), and Smooth Inverse Frequency (SIF).

GloVe embedding works on the word level. It uses a specific weighted least square model that uses a global word-word co-occurrence matrix for training to make efficient use of statistics. This co-occurrence matrix is constructed by training different corpora [14]. The GloVe vector of a sentence is the average of the GloVe vectors of its words.

USE works on the word, sentence, and paragraph levels where any word, sentence, or paragraph is converted into a vector of 512 dimensions. This encoder uses two different models. One makes use of the transformer architecture. This model achieves higher accuracy with greater model complexity and resource consumption. The other model is implemented as a Deep Average Network (DAN) which achieves efficient inference with slightly reduced accuracy [15].

SIF provides a new simple sentence embedding technique. This technique computes the weighted average of the word vectors in the sentence and then removes the projections of the average vectors on their first singular vector. The weight of a word w is *a / (a + p(w))* with *a* being a parameter and *p(w)* the estimated word frequency [16].

IV. Proposed Model

Our proposed model focuses on linear cognitive scripts in which events are provided in the form of “subject/verb/object”. For example, “audience enters stadium”, “audience thanks lecturer”, or “waiter delivers menu”. Events can also be provided in the form of “subject/verb”. For example, “movie starts”, “game ends”, or “lecture ends”. Finally, events can be in the form of “subject/adjective”. For example, “light on”, “light off” or “audience excited”. The subject and object of an event represent the event parameters and the verb or adjective represents the event action.

Before applying the evaluation model, some data processing must be executed. All scripts are converted into texts with every event converted into a sentence. For every script text, the average GloVe similarity is computed and stored. This is attained by computing the GloVe similarity of every sentence of the script with the remaining sentences of the script and then averaging these computed similarities. The pseudo code to compute he GloVe similarity between a sentence s and a sequence of sentences ss is shown in Fig. 2.

An important note to be considered is the assumptions that opposite events don’t exist in direct sequence in most cases. Opposite events are these events that have the same parameters but have opposite actions. For example, “light on” and “light off”, “audience enters stadium” and “audience leaves stadium”, and “movie starts” and “movie ends” don’t exist in direct sequence in most cases. All sequences of the blended script having subsequent opposite events are rejected when specifying the optimal positions of the events added to the target script. All opposite actions are stored before applying the evaluation model.

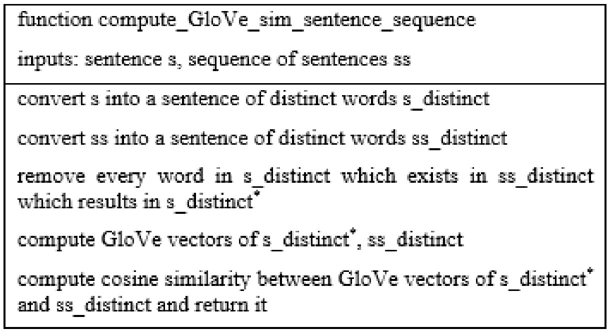
A. First Evaluation Phase

The first evaluation phase evaluates if the events selected from the base script can be added to the target script without changing its context. Since the GloVe embedding uses a global word-word co-occurrence matrix, it can be used to evaluate the probability of the existence of a sentence in the context of other sentences. When an event is selected from the base script, it is converted into a sentence and the GloVe similarity between this sentence and the target script text is computed and then divided by the average GloVe similarity of the target script text. The resulting value is defined as the GloVe similarity ratio between the event and the target script. If this ratio exceeds or equals a specified threshold, the selected event can be added to the target script.

B. Second Evaluation Phase

The second evaluation phase is used to specify the optimal positions to insert the events transferred from the base script into the target script. This can be done using text coherence evaluation techniques. The target script is converted into a text and the transferred events are converted into sentences. The optimal positions of the transferred events achieve the highest text coherence.

Text coherence evaluation techniques relying on the syntactic view of text coherence have a serious drawback when applied to cognitive scripts used in this work. This drawback is the reliance on the entities of the text while ignoring verbs and adjectives. Verbs and adjectives are very important in cognitive scripts because they represent the events’ actions while entities represent the events’ parameters. For example, consider the two events in the cinema script; “light on” and “light off”. These sentences have the same entity “light”. The entity’s role in the two sentences is a subject. If these events are exchanged, the coherence of the script text will not change although the logic of the text is completely different. The same problem occurs with events such as, “audience enters theater” and “audience leaves theater”.



```
function compute_GloVe_sim_sentence_sequence
inputs: sentence s, sequence of sentences ss

convert s into a sentence of distinct words s_distinct
convert ss into a sentence of distinct words ss_distinct
remove every word in s_distinct which exists in ss_distinct
which results in s_distinct*
compute GloVe vectors of s_distinct*, ss_distinct
compute cosine similarity between GloVe vectors of s_distinct*
and ss_distinct and return it
```

Fig. 2. Pseudo Code to Compute the GloVe Similarity between a Sentence and a Sequence of Sentences.

The technique used in this work relies on the semantic view of text coherence and uses (1) where the similarity between any two subsequent sentences is the cosine similarity between their corresponding embedding vectors.

V. Results and Discussion

A. Dataset

To the extent of our knowledge, there is no benchmark dataset for cognitive scripts. Four linear cognitive scripts adopted from [6] are used in this experiment; stadium, lecture, restaurant, and cinema. They were chosen because they provide a good variation from less detailed scripts such as stadium and lecture scripts to more detailed scripts such as restaurant and cinema scripts. The four scripts as converted into texts are shown in Fig. 3, Fig. 4, Fig. 5, and Fig. 6. In these figures, every sentence of the text represents an event of the corresponding cognitive script.

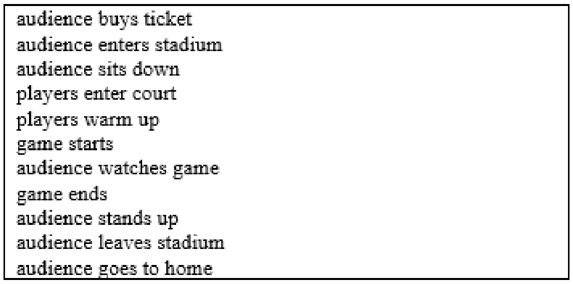
These four cognitive scripts are used to create a dataset to compute the GloVe similarity ratio threshold, evaluate the performance of the first evaluation phase, and evaluate the performance of the second evaluation phase. The procedure for creating the dataset is explained as follows:

- Every script of length n is split into all possible partitions of lengths n-1 and 1. The “n-1” events partition will represent a target script and the other partition will represent an event selected to be transferred to this target script.
- Again, every script of length n is split into all possible partitions of lengths n-2 and 2. The “n-2” events partition will represent a target script and the other partition will represent two events selected to be transferred to the target script.

Since this work focuses on enhancing the social dynamics of one cognitive script by transferring events from a contextually similar cognitive script, it is assumed that the core of the target script exists and at most two events are transferred from the base script to this target script.

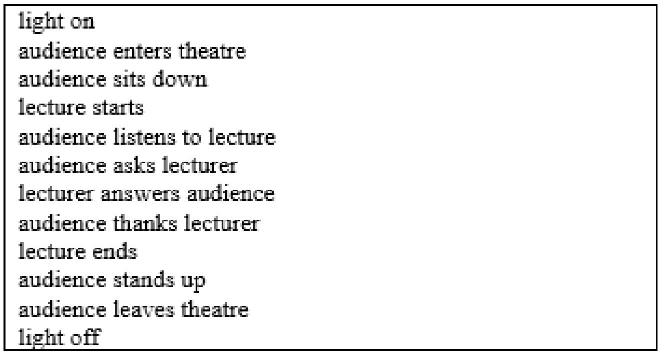
For a script of length n, the number of all possible partitions of lengths n-1 and 1 is n. Similarly, the number of all possible partitions of lengths n-2 and 2 is C^n_2. The four scrips will create a dataset of 340 script instances; 50 “n-1/1” script instances and 290 “n-2/2” script instances.

It is worth noting that “n-1” and “n-2” scripts are converted into texts and their average GloVe similarities are computed and stored.



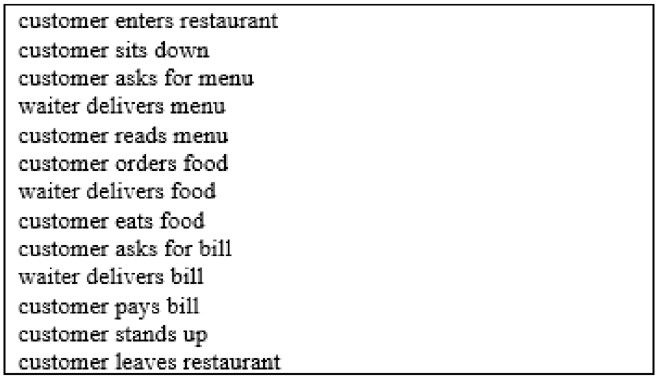
audience buys ticket
audience enters stadium
audience sits down
players enter court
players warm up
game starts
audience watches game
game ends
audience stands up
audience leaves stadium
audience goes to home

Fig. 3. The Stadium Linear Cognitive Script as Converted into Text.



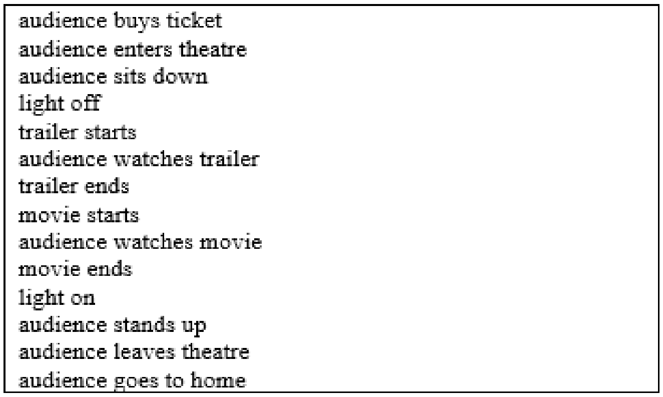
light on
audience enters theatre
audience sits down
lecture starts
audience listens to lecture
audience asks lecturer
lecturer answers audience
audience thanks lecturer
lecture ends
audience stands up
audience leaves theatre
light off

Fig. 4. The Lecture Linear Cognitive Script Converted into Text.



customer enters restaurant
customer sits down
customer asks for menu
waiter delivers menu
customer reads menu
customer orders food
waiter delivers food
customer eats food
customer asks for bill
waiter delivers bill
customer pays bill
customer stands up
customer leaves restaurant

Fig. 5. The Restaurant Linear Cognitive Script Converted into Text.



audience buys ticket
audience enters theatre
audience sits down
light off
trailer starts
audience watches trailer
trailer ends
movie starts
audience watches movie
movie ends
light on
audience stands up
audience leaves theatre
audience goes to home

Fig. 6. The Cinema Linear Cognitive Script Converted into Text.

B. Computing the GloVe Similarity Ratio Threshold

The proposed model uses GloVe vectors of 300 dimensions that are created by training Common Crawl (840B tokens, 2.2 M vocab, cased, 2.03 GB download). These vectors can be downloaded from https://github.com/stanfordnlp/GloVe.

For all “n-1/1” instances, the GloVe similarity ratio between the event and the “n-1” script is computed. For all “n-2/2” instances, the GloVe similarity ratio between every event of the two events and the “n-2” script is computed. The GloVe similarity ratio of some events can’t be computed since the parameters and actions of these events already exist in the target scripts. Examples of such events are “trailer starts”, “audience watches movie”, and “movie ends” when added to the cinema script.

Six different actions are chosen to be added to the “n-1” and “n-2” target scripts. Since the approach used for selecting events to be transferred to the target script implies parameter mapping [6], only the GloVe similarity ratios between these actions and the target scrips are computed. The selected actions are “sleeps”, “eats”, “listens’, “teaches”, “sings”, and “dances”. If the action exists in one of the target scripts, its GloVe similarity ratio will not be computed.

Human intervention is essential to decide for every action if it can be added or not to the target script. This human intervention focuses on cases where there is a certainty about accepting or rejecting the action. Cases, where there is an uncertainty about acceptance or rejection, are excluded. This is shown in Table I.

The script name which is underlined indicates that the action already exists in the corresponding script and its GloVe similarity ratio will not computed. From Table I, there is an uncertainty about some actions whether they are accepted or rejected in some target scripts. For example, there is an uncertainty about adding the action “listens” to the restaurant script since people may listen to music while eating. Similarly, people may eat while watching a game in the stadium or watching a movie in the cinema. The GloVe similarity ratios between the six actions and all “n-1” and “n-2” scrips are computed. These results are combined with the results of computing the GloVe similarity ratios using the “n-1/1” and “n-2/2” scripts and are used to specify the GloVe similarity ratio threshold which is empirically set to 0.8.

C. Evaluation of the First Evaluation Phase

The GloVe similarity ratio threshold specified empirically above is used to deduce the confusion matrix for this phase, which is shown in Table II, where True Positive (TP) = 507, True Negative (TN) = 1392, False Positive (FP) = 139, and False Negative (FN) = 60. Performance of this phase will be measured in terms of sensitivity, specificity, and accuracy. Sensitivity is calculated as *TP * 100 / (TP + FN)*, specificity is calculated as *TN * 100 / (TN + FP)*, and accuracy is calculated as (*TP + TN) * 100 / (TP + TN + FP + FN)*. Sensitivity = 89.42%, specificity = 90.92%, and accuracy = 90.51%.

TABLE. I. ACCEPTANCE/REJECTION OF THE SIX ACTIONS WITH RESPECT TO THE FOUR SCRIPTS

Action	must be accepted in	must be rejected in
sleeps		Stadium, lecture, restaurant, and cinema
eats	Restaurant	Lecture
listens	Lecture	Stadium and cinema
teaches	Lecture	Stadium, restaurant, and cinema
sings		Stadium, lecture, restaurant, and cinema
dances		Stadium, lecture, restaurant, and cinema

TABLE. II. CONFUSION MATRIX OF THE FIRST EVALUATION PHASE

	GloVe similarity ratio >= 0.8	GloVe similarity ratio < 0.8
The event must be added to the target script	507	60
The event must not be added to the target script	139	1392

D. Evaluation of the Second Evaluation Phase

The transferred events are inserted into all possible positions in the target script. The coherences of all these texts are computed. The optimal positions of the transferred events are these positions that result in the highest coherence of the blended script text. The accuracy is defined as the percentage of the blended script texts rather than the optimal blended text that have coherence lower than to that of the optimal blended text.

Three sentence embeddings will be used and compared; GloVe, USE, and SIF. USE will be implemented by a light weighted version of the transformer model which can be used with limited computation resources but still gives good performance. This can be viewed at https://tfhub.dev /google/universal-sentence-encoder-lite/2. The code for computing the cosine similarity between two SIF sentence embeddings can be viewed at https://www.kaggle.com/ procode/sif-embeddings-got-69-accuracy.

To evaluate this phase, two cases will be tested:

- For all “n-1/1” instances, the event is added to the “n-1” target script.
- For all “n-2/2” instances, the two events are added to the “n-2” target script simultaneously.

Adding one event in all possible positions in an “n-1” target script will result in n blended texts. One of them is the optimal text. Accuracies are computed for all “n-1/1” instances of every script and then averaged for every one of the three sentence embeddings used. The accuracies of this case are shown in Table III and Fig. 7. The accuracy of every sentence embedding technique is the average of its accuracies for the four scripts. The accuracy for adding one event to a target script of length n-1 using GloVe is 86.52%, USE is 75.51%, and SIF is 72.23%.

TABLE. III. ACCURACIES OF ADDING ONE EVENT TO “N-1” TARGET SCRIPTS

Script	Accuracy of GloVe%	Accuracy of USE%	Accuracy of SIF%
Stadium	80.99	85.95	71.07
Lecture	88.19	70.14	59.72
Restaurant	84.02	63.31	81.07
Cinema	92.86	82.65	77.04

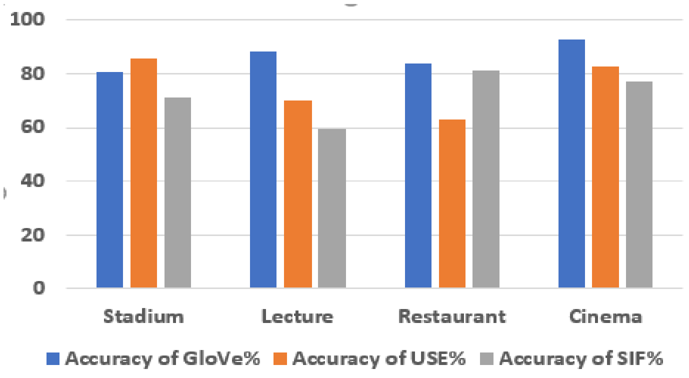


Fig. 7. Accuracies of Adding One Event to “n-1” Target Scripts.

Adding two events simultaneously in all possible positions in an “n-2” target script will result in P^n_2 blended texts. One of them is the optimal text. Accuracies are computed for all “n-2/2” instances of every script and then averaged for every one of the three sentence embeddings used. The accuracies of this case are shown in Table IV and Fig. 8. The accuracy of every sentence embedding technique is the average of its accuracies for the four scripts. The accuracy for adding two events to a target script of length n-2 using GloVe is 92.54%, USE is 79.02%, and SIF is 74.75%.

The GloVe embedding achieves the highest accuracy for the two cases. The accuracy of the second evaluation phase is the average of the accuracies of the two cases using GloVe embedding. The accuracy of the second evaluation phase is 89.53%. The overall accuracy of the evaluation model is the product of the accuracies of the first evaluation phase and the second evaluation phase. The overall accuracy of the evaluation model is 81.03%.

Another interesting metric to evaluate the proposed model is to study the effect of exchanging two events that share the same parameters but have different actions. In entity-based text coherence evaluation techniques, exchanging such events does not change the coherence of the script text although the logic of the script is completely different. The four scripts used in this work contains 14 event pairs of such type. They were exchanged such that only two events are exchanged per one time resulting in 14 different scripts. Coherences of these 14 scripts’ texts were computed and then compared to the four scripts’ optimal script texts’ coherences using the three sentence embeddings used in this work. For Glove embedding, 9 of these scripts had coherence lower than that of the optimal scripts. For USE, 5 of these scripts had coherence lower than that of the optimal scripts. For SIF embedding, 3 of these scripts have coherence lower than that of the optimal scripts. GloVe embedding achieved higher accuracy over USE and SIF in the case of exchanging two events that share the same parameters but have different actions.

E. Discussion

The evaluation model achieves a promising accuracy but with a high computational cost.

Transferring two events to a target script of length n-2 simultaneously requires P^n_2 i.e. n*(n-1) computations of text coherence to decide on the optimal blended script while, transferring them sequentially requires n-1 computations for the first event and n computations for the second event with a total number of 2n-1 computations. Future work should focus on reducing the computational cost and enhancing the accuracy.

TABLE. IV. ACCURACIES OF ADDING TWO EVENTS TO” N-2” TARGET SCRIPTS SIMULTANOUSLY

Script	Accuracy of GloVe%	Accuracy of USE%	Accuracy of SIF%
Stadium	88.76	91.54	69.06
Lecture	93.5	75.83	57.7
Restaurant	90.87	62.44	92.51
Cinema	97.04	86.25	79.72

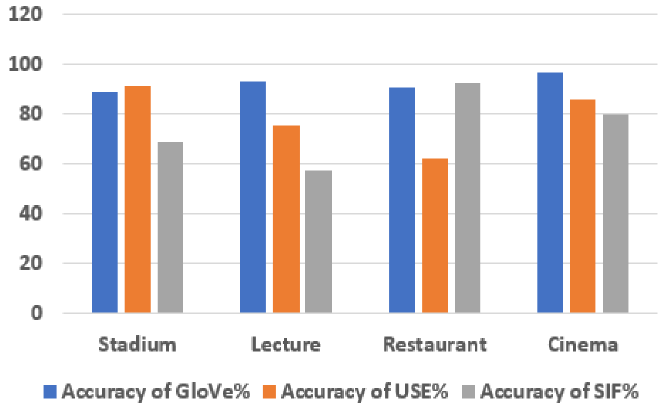


Fig. 8. Accuracies of Adding Two Events to “n-2" Target Scripts Simultanously.

The usage of a global word-word co-occurrence matrix can help explaining the reasons behind the GloVe embedding being successful in the first and second evaluation phases.

The accuracies of the four scripts using the Glove embedding ordered from the highest to lowest are that of cinema, lecture, restaurant, and stadium scripts. An explanation for the cinema script to be on top of the list is the fact that it has a lot of details and 5 pairs of start/end events. These events are “audience enters theatre” and “audience leaves theatre”, “audience sits down” and “audience stands up”, “light on” and “light off”, “trailer starts” and “trailer ends”, and “movie starts” and “movie ends”. Although the restaurant script has more details than the lecture script, the lecture script has more start/end pair events than the restaurant script has. This can explain why the accuracy of the lecture script is higher than that of the restaurant script. In general, two factors, that seem to affect the accuracy of a script using GloVe embedding, are details included in the script and the number of start/end pair events the script contains.

VI. Conclusion

This paper introduces a computational model for evaluating blended cognitive scripts resulting from transferring new events to a target script from another cognitive script known as the base script. The proposed model focuses on linear cognitive scripts consisting of events in the form of “subject/verb/object”, “subject/verb”, or “subject/adjective”. Subjects and objects are called the event parameters, while verbs or adjectives are called the event actions. Before an event is transferred from a base script to a target script, its parameters are mapped.

Four scripts are adopted and used to create a dataset of 340 script instances.

The evaluation process consists of two phases. The first evaluation phase evaluates if the selected events can be added to the context of the target script. The target script is converted into a text with every event converted into a sentence and the selected events are converted into sentences. The GloVe similarity ratio between every selected event and the target script is computed. If this ratio exceeds or equals a specified threshold, the event is transferred to the target script. The

GloVe similarity ratio threshold was empirically computed to be 0.8. The first evaluation phase achieved an accuracy of 90.51%.

The second evaluation phase is used to specify the optimal positions to insert the transferred events into the target script. The idea of the second evaluation phase is that the optimal positions of the transferred events are the positions that achieve the highest coherence of the blended script text. The text coherence evaluation technique used relies on the semantic view of text coherence where every sentence of the text is converted into an embedding vector, the similarities between every two subsequent sentences are computed as the cosine similarity of their embedding vectors, and then these similarities are averaged to compute the text coherence. Three sentence embeddings are used and compared. The GloVe embedding achieved higher accuracy over USE and SIF embeddings. The accuracy of the second evaluation phase is 89.53% using GloVe embedding.

The proposed model achieved an overall promising accuracy of 81.03% but with a high computational cost.

VII. Future Work

Future work will focus on reducing the computational cost and enhancing the accuracy. Different approaches may be suggested. One approach is to test the addition of two events sequentially firstly, without any precedence of one event over the other and secondly, with applying a precedence rule in transferring the two events such as transferring the event with the higher GloVe similarity ratio first.

Another approach is to convert the second evaluation phase into an optimization problem. The second evaluation phase searches for the optimal positions to insert the transferred events. Optimization techniques such as Genetic Algorithms (GA) or Particle Swarm Optimization (PSO) may be used.

With the two previously mentioned approaches, the text coherence of the blended script text may be measured using deep learning techniques such as using a Convolutional Neural Network (CNN) [17].

Converting the second evaluation phase into a problem of sentence ordering may be a solution. Recurrent Neural Networks (RNNs) may be used in this approach [18].

References

[1] M. Gawish, S Abbas, M. G. Mostafa, and A. B. M. Salem, “Learning cross-domain social knowledge from cognitive scripts,” Proceedings of the 8th International Conference on Computer Engineering & Systems (ICCES), IEEE, 2013.

[2] D. Gartner and L Smith, Analogical reasoning, Encyclopedia of Human Behavior, 2nd ed., Elsevier, pp. 130–136, 2012.

[3] G. Fauconneit and M. Turner, "Conceptual integration networks,” Cognitive Sciences, vol. 22, no. 2, pp. 133–268, March 1998.

[4] R. Hodhod and B. Magerko, “Pharaoh: conceptual blending of cognitive scripts for computationally creative agents,” Proceedings of the Twenty-Seventh International FLAIRS Conference, May 2014.

[5] J. Permar and B. Magerk, “A conceptual blending approach to the generation of cognitive scripts for interactive narrative,” Proceedings of AIIDE, 2014.

[6] M. Gawish, Developing an intelligent computation model for human episodic learning, Master thesis, Ain Shams University, May 2017.

[7] M. Lapata and R. Barzilay, “Automatic evaluation of text coherence: models and representations,” Proceedings of the 19th International Joint Conference on Artificial intelligence, pp. 1085–1090, 2005.

[8] A. B. M. Salem and M. Gawish, “Study on analogical reasoning methodologies for developing analogical learning systems,” International Journal of Circuits and Engineering, vol. 1, pp. 54–61, 2016.

[9] R. Hodhod, B. Magerko, and M. Gawish, “Pharaoh: context-based structural retrieval of cognitive scripts,” International Journal of Information Retrieval Research (IJIRR), vol. 2, no. 3, pp. 58–71, 2012.

[10] T. Veale and D. O’Donoghue, “Computation and blending,” Cognitive Linguistics, vol. 11, no. 3-4, pp. 253–281, 2000.

[11] T. Pederson, S. Patwardhan, and J. Michelizzi, “WordNet:: similarity: measuring the relatedness of concepts,” Demonstration Papers at HLT-NAACL, pp. 38–41, 2004.

[12] H. Liu and P. Singh, “ConceptNet—a practical commonsense reasoning tool-kit,” BT technology Journal, vol. 22, no.4, pp. 211–226, 2004.

[13] P. Foltz, W. Kintsch, and T.K. Landauer, “The measurement of textual coherence using latent semantic analysis,” Discourse Processes, vol. 25, no. 2-3, pp. 285–307, 1998.

[14] Pennington, R. Socher, and C. D. Manning, “GloVe: global vectors for word representation,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, 2014.

[15] D. Cer et al., “Universal sentence encoder,” ArXiv, 1803.11175, 2018.

[16] S. Arora, Y. Liang, and T. Ma, “A simple but tough-to-beat baseline for sentence embeddings,” Proceedings of ICLR, 2017.

[17] B. Cui, Y. Li, Y. Zhang, and Z. Zhang, “Text coherence analysis based on deep neural network”, Proceedings of ACM on Conf. Info. Know. Manag. (CIKM), pp. 2027 – 2030, Nov. 2017.

[18] L. Logeswaran, H. Lee, and D. Radev, “Sentence ordering and coherence modeling using recurrent neural networks”, Proceedings of AAAI, 2018.

Systematic Literature Review of Identifying Issues in Software Cost Estimation Techniques

Muhammad Asif Saleem[1], Tahir Alyas[3], Asfandayar[5]
Department of Computer Sciences
Lahore Garrison University
Lahore, Pakistan

Rehan Ahmad[2], Asif Farooq[6], Kahawaja Ali[8]
Department of Computer Science
The University of Lahore, Lahore, Pakistan

Muhammad Idrees[4]
Department of Computer Sciences & Engineering
The University of Engineering and Technology
Narowal Campus, Pakistan

Adnan Shahid Khan[7]
Faculty of Computer Sciences and Information Technology
Universiti Malaysia Sarawak, Sarawak, Malaysia

***Abstract*—Software cost estimation plays a vital role in software project management. It is a process of predicting the effort and cost in terms of money and staff required for developing the software system. It is very much clear that software project will be successful if its estimated cost will be near to the real cost. When the project is at the acquisition stage the least details are available about a software project to be developed, due to which Problems arises in cost estimation. As the stages move on details increases for software development of software which is quite fruitful in cost estimating. However, it can be considered that estimating the software cost in the first phases will produce better results. In this research cost estimation techniques are discussed along with the issues in that particular technique and focus will be on understating the points or issues which cause hurdles or issues in estimating the cost of the software project.**

Keywords*—*Cost estimation; COCOMO; qualitative; use case point, PCA

I. Introduction

Estimating the software cost is very important but it is difficult to do so as well. In the early era of software development, there were very fewer instructions in software, as the software size is increased the accuracy in the cost of estimation also increased. Nowadays large software projects increased up to 25 million lines of source code [1]. Which may need 1000 software developers to code which may take 5 years to complete. The cost of that project may increase to 400 million dollars. So mistakes in cost estimation of these types of project will be very ample serious indeed. A huge amount of large software systems could not meet the deadlines, runs over budget or face cancellation of the project due to underestimation or overestimation of software cost at the phase of requirement engineering. In fact, overconfidence in measuring the software cost is a vital cause of expansion in software budget, failures, and litigation. Now a day's software industry is imparting the main role in running every type of business in the public and private sector. It can be considered that the government of a state or a private corporation builds thousands of software applications and may change them according to their needs every year. So being a welcome agent of a software project in this world, every software house who dealing in large or complex software system development pays huge focus on estimating the cost of the project which is going to be developed. In this study different software cost estimation techniques proposed and finding their issues with the help of systematically selecting the researching databases, selecting article and deciding their selection or rejection criteria. This study will help the researchers as well as practitioners for selecting suitable software cost estimation model. For this systematic literature study article is divided into multiple sections which are as:

- Methodology for literature study
- Literature Study
- Software cost estimation techniques and Disadvantage
- Issues and problems in software cost estimation models
- Final Conclusion and future work

II. Methodology for Literature Study

The systematic literature review has been utilized to do this research. It is an appropriate and repeat procedure to record relevant points of interest in the exact research range for inspecting and examining all current research identified with research questions. Thus, this exploration consolidates following phases like Categories definition, Review protocol development, and selection and rejection criterion and Search process.

A. Category Definition

We have characterized six categories with a specific end goal to sort out the search result. This order will fundamentally manage the correctness of the exact responses. The brief description is given below.

B. General Category

There may be various studies in software cost estimation along with different modular approaches. (Expenses on modeling, transformation, verification, and simulation). As per our survey study, we have categorized the software cost estimation in the categories like Algorithmic approach, Non Algorithm approach, and Expert opinions.

TABLE. I. SEARCH PROCESS

	No. of Articles							
	Library	***Operator***	***Library***	***Operator***	***Library***	***Operator***	***Library***	***Operator***
Keywords	IEEE	AND	ACM	AND	SPRINGER	AND	ELSEVIER	AND
Software cost estimation	955	32	3154	29	327	3	46,358	2
Software Cost determination	97	2	3055	1	102	0	4,384	0
Software effort estimation	454	55	2774	37	432	10	5814	13
Software cost calculation	246	8	2764	0	330	0	1169	0
Software cost detection	1037	1	3,205	0	469	0	9140	0
Tools software cost	1586	1	3,246	1	868	0	2678	1
Framework software cost	1314	1	3,698	0	855	0	16206	0
Models Software Cost estimation	4670	12	5,051	0	1164	0	8989	3
Methods software cost	3169	4	4,577	1	1111	0	25534	0
Techniques software cost	2332	3	4,651	0	1086	0	21216	0
Software cost estimation steps	525	4	2,989	0	1009	0	7795	0
Software cost estimation phases	740	1	2,815	0	687	0	4,745	0
Total	17125	123	41979	68	8008	13	148213	19

C. *Review Protocol Development*

As the categories for software costing is defined our next phase is to develop a protocol for searching the relevant articles. For review protocols, we have considered search process, selection and rejection criteria, quality assessment and data extraction.

D. *Search Process*

As the table shows above, it shows that we have chosen four logical databases (i.e. IEEE, ELSEVIER, SPRINGER, and ACM) with a specific end goal to complete our research. These logical databases maximum result regarding our topic and we have selected the required article via the self-defined process described below.

We have selected year-wise range from "2013–2017" for selection of research articles from our selected logical databases. We use different keywords for selecting articles (e.g. SCE, SCDE), different operators like AND OR are also used for making our search accurate. We have applied AND operator in abstract, keywords and title of the article. In this way the result from logical databases are refined i.e. non-relevant articles are removed in this way. The complete search process is defined and practices given in Table I.

III. SELECTION CRITERIA

We characterize the solid paradigm for the choosing and rejecting of research works. Six parameters are characterized to guarantee the correctness of the appropriate responses of our research questions. The research work will be chosen on the premise of these parameters as given underneath.

A. *Subject-Relevant*

Select the research work just if that it comes fits our research settings made by our category design. It must encourage the appropriate responses of our research addresses and should be applicable to one of the predefined categorizations. Dismiss insignificantly explores those don't have a place with any of the predefined categories.

B. *Year Wise Range 2013-2017*

Chosen research work must be distributed from 2013 to 2017 by inserting filter an off year-wise selection in every database which will result reduce overall results and helps in finalizing the research article. Reject all the article less our selection criteria set in year wise selection mechanism.

C. *Publisher*

Selected research work must be published in one of the four renowned scientific databases i.e. IEEE (IEEE Scientific Database, 2014), SPRINGER (Springer, 2014), ELSEVIER (Elsevier, 2014) and ACM (ACM, 2014).

D. *Crucial-Effects*

Chosen research work must have vital beneficial results with respect to software cost estimation under software project management approach. Dismiss the research work if its proposal does not have a large contribution estimation of software cost.

E. *Results-Oriented*

Selected research work must be results-oriented means producing a result, not a simple survey. Have a brief look on the result section of the selected article and verify that that contribution of this article has major worth in the field of software cost estimation. Result verification must be carried out by a powerful survey if it does not so dismiss the work.

IV. REJECTION CRITERIA

A. *Repetition*

All the research in a specific research setting can't be incorporated. Thus, dismiss the research if these are indistinguishable in the given research setting and just a single of them is chosen.

B. Rejection on the Title Bases

Selected research work can be justified by having a brief look at the title of the research articles. It may need some expertise while judging the article but it will provide a fruitful result. The proposition and extreme results of the research must be upheld by strong certainties and experimentation. Dismiss the work if its title is not matching you're related to the research topics.

C. Rejection on Basis of Abstract

Sometimes it is very much hard to take decision while selecting the research article by checking with the title of the article so in this regard you should read the abstract of the article from which you can get proper information regarding the article.

V. Literature Study

Software costing is important to phase in software project management [1]. It is also crucial and difficult for Development Company and client as well. Software cost estimation is essential to the contract, negotiating with the client and writing for the request of the proposal, Monitoring and controlling. Overestimation in cost may result in maximum resources promised to the project or bidding in contract, due to all these issues project may not win for the company. Not winning the project may lead to termination to the job. Accurate cost estimation is important because:

- as per the overall plan of business, it may be helpful in the classification and prioritization of development project [2];
- accurate cost estimation may be helpful in calculating the resources for committing to the project;
- for the impact of change and support preplanning;
- it is better to manage software project when resources areas the needs; and
- customers are always in view that the real cost of software should be equal to estimated costs.

Measuring software cost depends upon on from these factors:

- Software Effort (Time, resources)
- Duration of project (as per calendar time)
- Software Cost (Amount)

Experts have studied these three problems given below:

- Which model of software cost estimation is to use?
- Which software size measurement to use lines of code (LOC), function points (FP), or feature point?
- What is a good estimate?

VI. Expert Judgment

In this method, we consult with an expert in software cost estimations and a group of experts who have experience in this field [3, 4]. By using their expertise and by understanding the nature of the project we came at the point to estimate the cost of the project.

This method has several weak points which are listed below:

- Non-quantifiable.
- Documenting the factor taken from an expert is difficult.
- Maybe the opinion of experts be biased [5].
- As the total information is not available at the initial phase so the proposed estimation may lead to the wrong [6].
- This process may be time-consuming because of lots of discussion among experts [7].
- Due to personal limitation or time may this method not be feasible.

VII. Algorithmic

In the algorithmic cost estimation methods, mathematical approaches are used for the purpose of estimation [8]. These mathematical approaches are based on historical data and research, which are used as inputs in function points, source lines of codes and other cost drivers. The algorithmic models are studies on large scale like the Putnam Model, COCOMO model and function point models.

- Issues in the algorithmic model are as follows:
- In this method dealing with exceptional conditions [5].
- False cost driver ratings and poor sizing inputs will produce poor costing results [9].
- Some factors may not be quantified resulting in bad costing results.

VIII. COCOMO

COCOMO Model is known as cost constructive model [10]. This model is widely used model in branch of cost estimation of software. Basic COCOMO Model has very simple form which can be described as:

MAN+MONTH = KI *(delivered source instructions) ^{K2}.

Here k1 and K2 are the parameters which are completely dependent on the environment of development and application under process. Estimation in the COCOMO model can be more precise if the following factors are kept in mind while following the steps of COCOMO models [2]. Characteristics of software under development, Qualification, and experiences of team members, Environment to be used for development purpose. Some of these factors are:

- Complexity of Software
- Reliability of software
- Size of database
- Efficiency required

- Teams experiences involved in the development
- The capability of developers involved in the analysis process
- Experience of software team with respect to computer and programming languages which are going to use in development purpose.

Many of the above issues may affect the MAN and MONTH required in order. COCOMO model can perform well if the requirement of software to be developed might be clear and stable. Which is mostly not clear or stable due to which its efficiency in these types of projects is not up to marks. This model is the regression model as it is based on 63 selected papers [9]. Primary inputs for this model is KDSI. Problems and issues in these models are described as:

- At the initial phase, there are lots of uncertainty in cost estimation so it is not possible to get near to the exact value of cost.
- COCOMO model is resulted by having the analysis of 63 selected papers. It usually has some problems at environment due to which recalibration is mandatory.

As per Kemmerer's research, the average bug rate in maximum versions of the model is recorded 601% [11]. The first version of the COCOMO model was developed in 1981. Nowadays it is facing problems in estimating the cost of software developed to new life cycle processes and capabilities including rapid-development process model, reuse-driven approaches, object-oriented approaches, and software process maturity initiative. Due to all these issues faced in this model new approach is designed which is named as COCOMO II.

Disadvantages in this models are described as follows:

- This model completely forgets requirement and all documentation [12]
- It also forgets customer's skills like knowledge, cooperation, and other parameters
- It didn't focus on the impact of security and safety
- It ignores personnel turnover level
- It doesn't consider hardware issues
- It is dependent on time spent in every phase.

IX. Putnam Model

This model is also algorithmic model developed after COCOMO II [13]. It is very much popular model with respect to cost estimation.

Disadvantages of this model are given as under.

This model is dependent on the size of the software to develop by knowing it or being able to estimate it correct. As we know there is a factor to great uncertainty in knowing the size of the software. It is understood that if the size of the software is not accurate than the cost estimated by the Putnam model will not be correct. As per Kemmerer research Putnam model is based on SLIM which is 773.00% [1].

X. Function Point Methods

Function point methods method was created by Allan Albrecht at IBM and this model published in 1979 [10, 14]. This model has many advantages over source line of code method count of sizing. The function point data collection has two major motives firstly as per the desire of the manager to check out the level of productivity. Secondly, it is used in the estimation of cost.

Problems or issues in this model are briefly described as [3]:

- This model works on the subjective approach with having maximum judgments considered.
- Lots of models are like effort or cost estimation is based on LOC so this should also be converted.
- Amount of data research is very much less in this model.

XI. Analogy Based

Estimation in this analogy based model is carried by comparing the proposed project with already developed projects with same requirement [15]. For this purpose, data is extracted from the already extracted project and compared with the proposed project. This methodology can be used on the project level as well as a component level [16].

Estimating by analogy is straight forward [2]. it is a very efficient way of expert judgment when experts were often in search for analogous situations so as to inform their opinion [17].

Disadvantages in these methods are as follows:

- Selection of projects for getting data from already developed is quite a difficult task [18].
- We are deriving an estimate for the new project with the help of already developed values from the randomly selected projects. Possible situations may involve means and weighted means which may provide more impact on the near to analogies.

XII. Top-Down Methods

The top-down strategy is also known as a macro method. In this method, the project cost is estimated from the universal properties of the project [19]. Then the project is divided into sub-projects or software component. This method can be used at the early phase of development because at that stage minimum derails are available which can be tackled in this model easily [20]. This method is useful when no detailed information is there.

Disadvantages of these methods are as follows:

- This model mostly ignores the low-level problems in software cost estimation that may affect the estimation of software cost [21].

- This model is unable to provide a detailed basis of decisions or estimations.

XIII. Bottom-Up

In the bottom up [22] approach every single component's cost is measured and then at the end total cost is measured by calculating every component' s cost. The most leading model which is using this approach is COCOMO [19, 23].

The problems or issues in these methods are described as follows:

- Cost estimated in this method may not be correct because most of the details are not available at early phases [24].
- This method is time-consuming.
- This method may not be able to use when the resources are limited.

XIV. Disscussion

Costing for software projects have distinctive elements that make the cost estimation so hard [25]. The complex ranges that impact the cost estimation are:

- Lacking in abilities of Software and maximum utilization of already available skills.
- Optimum utilization of COTS parts.
- Estimation of software.
- System of frameworks, specifically security issues.
- Interoperability.
- Catching of Requirements.
- Obsolescence.
- Designing software.

Estimating Software cost is so difficult in light of the fact that:

- The software truly is substantially difficult. Consider the number of requests for greatness included. Time scale 24x7 down to 100 nanosecond clock cycles is a LOT of requests of extent: beyond any reasonable amount to completely fathom. Henceforth the innate trouble. Consequently the vulnerability.
- Because you have a point where the project should leave meanwhile client is unable to understand how the clients fulfill its need [26].
- The major fault at client side that the client does not focus on "what he needs" More awful, they can't know until the point when they begin to work with the arrangement innovation which at that point changes the idea of the issue [26].
- We focus on the upcoming by taking a gander at the past. So I think there are two principle issues with estimation [27].
- If we were made a request to finish the something we have never done, we can't know to what degree, it will take since we have never performed it sometime in the near future.
- If we do not manage a record of what extent matter may take place, even when we're made a request to accomplish something comparable once more, we won't have the information whereupon to gauge.

XV. Conclusion

From this system literature study following conclusion are drawn:

- The cost estimation is a complex task and all the methods proposed have multiple advantages and disadvantages. Like some models performs better in large projects, some performs better on small projects and some performs better on global software development.
- None of the factors, which may affects cost estimation be ignored while estimation cost of software.
- For best and consistent cost estimation, the predictive attributes from the data set may considered. Backward input approach maybe use for selection these attributes.
- In this study, multiple methods are considered, the best efficiency is provided by collaborative methods.
- For better results, Experts opinion must be kept in mind. Biasness of experts may judge and removed.

The future work in cost estimation is proposing a novel model using hybrid approach. Which will be further enhanced by expert's opinion.

References

[1] P. Agrawal and S. Kumar, "Early phase software effort estimation model," in Colossal Data Analysis and Networking (CDAN), Symposium on, 2016, pp. 1-8.

[2] S. Bilgaiyan, S. Mishra, and M. Das, "A review of software cost estimation in agile software development using soft computing techniques," in Computational Intelligence and Networks (CINE), 2016 2nd International Conference on, 2016, pp. 112-117.

[3] W. Han, T.-B. Lu, H. Jiang, and W. Li, "A Study on the Improvement of Software Effort Estimation," in Control and Automation (CA), 2014 7th Conference on, 2014, pp. 24-27.

[4] M. El Bajta, A. Idri, J. L. Fernández-Alemán, J. N. Ros, and A. Toval, "Software cost estimation for global software development a systematic map and review study," in Evaluation of Novel Approaches to Software Engineering (ENASE), 2015 International Conference on, 2015, pp. 197-206.

[5] K. Usharani, V. V. Ananth, and D. Velmurugan, "A survey on software effort estimation," in Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on, 2016, pp. 505-509.

[6] M. Saroha and S. Sahu, "Software effort estimation using enhanced use case point model," in Computing, Communication & Automation (ICCCA), 2015 International Conference on, 2015, pp. 779-784.

[7] P. Pandey, "Analysis of the techniques for software cost estimation," in Advanced Computing and Communication Technologies (ACCT), 2013 Third International Conference on, 2013, pp. 16-19.

[8] L. V. Patil, R. M. Waghmode, S. Joshi, and V. Khanna, "Generic model of software cost estimation: A hybrid approach," in Advance Computing Conference (IACC), 2014 IEEE International, 2014, pp. 1379-1384.

[9] M. Phongpaibul and P. Aroonvatanaporn, "Standardized cost estimation in Thai government's software development projects," in Computer Science and Engineering Conference (ICSEC), 2015 International, 2015, pp. 1-6.

[10] S. Islam, B. B. Pathik, M. H. Khan, and M. M. Habib, "A novel tool for reducing time and cost at software test estimation: An use cases and functions based approach," in Industrial Engineering and Engineering Management (IEEM), 2014 IEEE International Conference on, 2014, pp. 312-316.

[11] N. Garcia-Diaz, C. Lopez-Martin, and A. Chavoya, "A Comparative Study of Two Fuzzy Logic Models for Software Development Effort Estimation," Procedia Technology, vol. 7, pp. 305-314, 2013.

[12] J. Huang, Y.-F. Li, and M. Xie, "An empirical analysis of data preprocessing for machine learning-based software cost estimation," Information and Software Technology, vol. 67, pp. 108-127, 2015.

[13] L. L. Minku and X. Yao, "Ensembles and locality: Insight on improving software effort estimation," Information and Software Technology, vol. 55, pp. 1512-1528, 2013.

[14] H. Rastogi, S. Dhankhar, and M. Kakkar, "A survey on software effort estimation techniques," in Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-, 2014, pp. 826-830.

[15] S. Garg and D. Gupta, "PCA based cost estimation model for agile software development projects," in Industrial Engineering and Operations Management (IEOM), 2015 International Conference on, 2015, pp. 1-7.

[16] D. Toka and O. Turetken, "Accuracy of Contemporary Parametric Software Estimation Models: A Comparative Analysis," pp. 313-316, 2013.

[17] N. Mittas, E. Papatheocharous, L. Angelis, and A. S. Andreou, "Integrating non-parametric models with linear components for producing software cost estimations," Journal of Systems and Software, vol. 99, pp. 120-134, 2015.

[18] Y.-S. Seo, D.-H. Bae, and R. Jeffery, "AREION: Software effort estimation based on multiple regressions with adaptive recursive data partitioning," Information and Software Technology, vol. 55, pp. 1710-1725, 2013.

[19] S. Al-Qudah, K. Meridji, and K. T. Al-Sarayreh, "A Comprehensive Survey of Software Development Cost Estimation Studies," pp. 1-5, 2015.

[20] A. Ren and C. Yun, "Research of Software Size Estimation Method," pp. 154-155, 2013.

[21] F. Sarro, A. Petrozziello, and M. Harman, "Multi-objective software effort estimation," pp. 619-630, 2016.

[22] D. Kashyap and A. K. Misra, "Software development cost estimation using similarity difference between software attributes," presented at the Proceedings of the 2013 International Conference on Information Systems and Design of Communication, Lisboa, Portugal, 2013.

[23] C. S. Stokes, A. R. Simpson, and H. R. Maier, "A computational software tool for the minimization of costs and greenhouse gas emissions associated with water distribution systems," Environmental Modelling & Software, vol. 69, pp. 452-467, 2015.

[24] A. Girasella and F. Pagin, "An UML-based approach to software development cost estimation," presented at the Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, Torino, Italy, 2014.

[25] M. Jørgensen, "Communication of software cost estimates," pp. 1-5, 2014.

[26] T. Urbanek, Z. Prokopova, R. Silhavy, and V. Vesela, "Prediction accuracy measurements as a fitness function for software effort estimation," Springerplus, vol. 4, p. 778, 2015.

[27] L. Kompella, "Advancement of Decision-Making in Agile Projects by Applying Logistic Regression on Estimates," pp. 11-17, 2013.

A Novel Student Clustering Model for the Learning Simplification in Educational Environments

Khalaf Khatatneh[1], Islam Khataleen[2], Rami Alshwaiyat[3], Mohammad Wedyan[4]
Computer Science Department, Al-Balqa Applied University, Al-Salt, Balqa, JORDAN

***Abstract*—Students' clustering is considered to be a method of granting students many different ways of being able to learn by taking into account the student's degree. The definition of clustering can be "a group of steps that divide a group of information or (things) in a group of valuable secondary classes, which are named clusters. This implies that a cluster can be a group of things that are "alike" among them and that can be "not alike" to things that are part of other types of clusters. Thus, in this study we want to make students into different clusters or groups in order to organize students and lecturers into more interactive ways, and to extract an education that is related to some particular information and to produce some comments and feedback to the academic teacher. Conclusion, this is the specific research effort that forms the final phase. In particular, it is the result of satisfaction when deviations exist in the artifact behaviour, which is derived from the (multiply) revised hypothetical predictions. Accordingly, the results are then declared as 'good enough'.**

Keywords—Clustering; students groups; education inter-active ways; students' clustering; student's degree; valuable secondary classes

I. Introduction

Each has its own information and data that are available in different departments. There are several demographic, educational and situational changes that are collected and processed for each involved student entering the school. Although schools have high amounts of information spread everywhere through their campuses, they are still facing some major issues on how proficient and efficient should education be within the teaching atmosphere. Additionally, technological developments have brought more efficient and large quantities of data and information. The use of knowledge is concealed through the majority of these quantities where there is a persistent need and necessity of detecting and making use of this grasp, for instance in student's information. The importance of understanding the educational system could involve several opinions and suggestions of students on different offered courses.

II. Related Work

Several studies have been carried out through learning for assessing teaching and learning, and for evolving novel methods as alternatives to old methods in teaching and learning. Old learning is mainly associated with replying to the questions or inquiries starting with "W" and "How" that concern a large portion of students, only in case of being carefully taught, group teachings by nature neglects the presumptive "which" and "why" inquiry, which makes a major concern on a small portion of students [1]. Moreover, old learning relies on the discussions of lectures in identical methods to the entire students. On the other hand, old learning relies on the grades and individualities of the students to evaluate the acquisition of knowledge [2]. Furthermore, every student takes identical data about the subjects. Several methods were proposed in learning, and which proved to be more efficient than the old learning, for example teaching through online instruments and students' clustering.

The data clustering of the students permits the to detect the students who are registering for certain academic courses including the students' identities where this fact could be beneficial for popularising particular goals, as well as from educational aspect. Teachers for particular subjects might also get data related on students to improve education.

Accordingly, this study aims at proposing an application of information mining methods to make information Student groups with their academic, demographic, and attitudinal changes to assist and estimate education. grouping algorithms was used in a big variety of implementations for instance image categorization and documentation restore, in education such as evaluate students based on incorrectness in practices and evaluate students dependent on their conduct of improving subject style [3] .Concerning attribute clustering and related attributes are distributed into categories depending on the relativeness among them.

A. Process and Tools for Tropos Development

In Tropos, there are five basic stages for software development. These stages are as follows: early-requirements analysis, late-requirements analysis, architectural design, detailed design, and the implementation stage. The analysis of early requirements is the first stage and its emphasis is on gaining an understanding of the current organisational setting where one will present the system. The analysis of late-requirements is the second stage and it focuses on analysing the system–to–be. The architectural design is the third stage and it identifies the system's global architecture in terms of the subsystems. The detailed design is the fourth stage and it is responsible for determining the system agents' micro level. Implementation is the last stage and its emphasis is on generating code based on the detailed design's characteristics.

Models serve as the basis of the software development process. These models include the design and the requirements. They also serve as the main artefacts. The conceptual modelling language is used to structure these models. As seen in Fig. 1, this activity is called Agent-Oriented (AO). The modelling language has several main concepts. These concepts are goal, actor, plan, and dependency in attaining the goal. The modelling tool of TAOM4e can be used to perform agent-oriented modelling. This tool is also capable of supporting automatic code generation based on the specifications of Jadex or Troposto JADE. It is able to do so by utilising the Tropos Meta model concepts and the concepts of target implementation languages.

Furthermore, other techniques are available for use in the tool-supported analysis in Tropos. One of these techniques is referred to as requirements validation, which is performed by checking the model or conducting a formal analysis on the system design and the goal models of requirements. Such analyses are specifically utilised in complex models.

The goal-oriented testing action is seen as one of the fundamental exercises to demonstrate AO and code age exercises. These experiments were specifically obtained from the AO details that helped formulate the procedure of advancement. Furthermore, the end goal of being able to assist in testing and approval was also kept in mind [4].

B. TAOM4e and Functions of Code Generation

Tropos can be modelled graphically using the TAOM4e framework, which supports the modelling throughout the complete stages of Tropos. As shown in Fig. 2, it is seen as a plug-in for the eclipse project and it also serves as an extension for the existing plug-ins. Illustrates how the EMF plug-in provides a code generation and modelling framework that can be used to develop tools and applications based on the model descriptions that are given in XMI.

The Tropos' meta-model is utilised at the top of EMF (for the TAOM4e model). The Graphical Editing Framework was then used to illustrate how the graphical representation of the model as well as the differences in prospective (TAOM4e platform). The Tefkat plugin can then be used to help turn top level plans and the decompositions of such plans to the UML's activity diagrams. Any UML2 editor can then be used to modify the resulting diagrams. The sequence diagrams are responsible for determining the communication protocols that are used among agents.

The TAOM4e modeller provides the TAOM4e generators so that code skeletons can be derived for the "JADE and Jadex" agent platforms. This derivation is performed from the goal model of Tropos or from the UML specification of the detailed design artefacts [5]. The TAOM4e generators involve Tropos-2UML, t2x, and UML2-JADE. The UML2-JADE is responsible for producing the JADE agent code based on the UML activity and sequence diagrams, where the Tropos plans are thoroughly provided. The latest version is capable of producing UML activities diagrams from the Tropos' goal model. The knowledge level refers to the portion where the agent is given the responsibility to select the correct plans in order for the goals to be accomplished. In this part, the knowledge level is made up of goals and decomposition, means–end relations to plans, and contributions and dependencies to other agents. The knowledge level inclusions serve as inputs for the t2x tool. This tool uses the architecture of BDI as a basis in producing skeletons for agents. One can then perform these skeletons on the Jade BDI agent's platform.

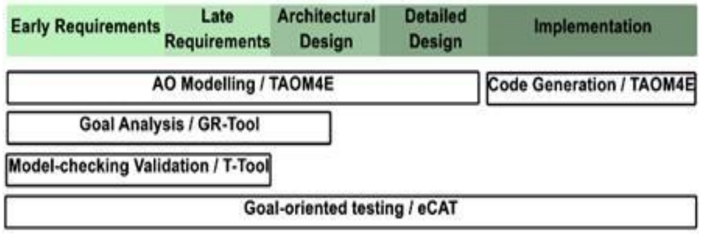


Fig. 1. Phases of the Development Process: Activities and Supporting Tools.

The skeleton code generated can then be applied to the software agent's reason ale. This includes agent definition file (ADF) that is presented in an XML format, and the identification of the beliefs, messages, goals, and plans of every system agent that is part of the GM [6]. One can then apply the single plans in the Java files, which are related to the ADF elements.

III. Test Description

A. Goal

Nomenclature lists (lists of symbols and definitions) generally follow the Abstract and Index Terms and precede the introduction.

B. Tested

This study measured the round-trip time, which can be defined as the time needed to conduct a circular ACL message exchange between a receiver agent and a sender agent.

To evaluate the platform scalability, the study will start with a single Sender/Receiver couple. The number of couples increased so that one could observe the manner by which the round-trip time increases. For each measurement, the average round-trip time (avgRTT) should ideally be obtained. This refers to the total measurement time divided by the amount of times that every couple exchanges a message as well as the ctual number of couples. Thus, avgRTT refers to the average time it takes to send a message and obtain a reply.

Fig. 3 shows that the Sender i-th only communicates with the Receiver i-th, where each couple exchanges 10000 messages. A seven characters string occupies the content field of each message.

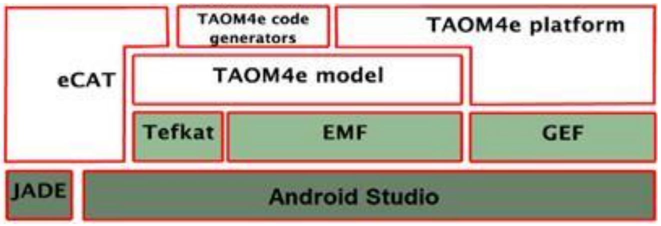


Fig. 2. TAOM4e and eCAT Architecture.

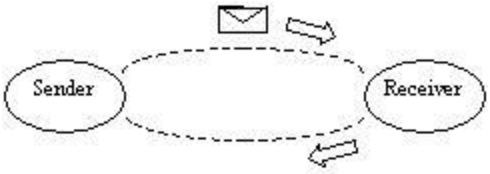


Fig. 3. ACL Round-Trip Message.

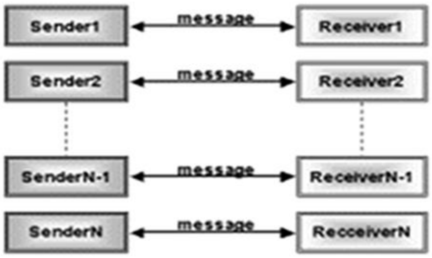


Fig. 4. Architecture of a Testbed Scenarios.

One can implement the two agents using the Cyclic Behaviour class (i.e. the JADE abstraction that is used for a repetitive cyclic agent task) to manage incoming messages. The following are the steps executed by both agents receiver and sender shown in Fig. 4.

The initiator role is filled by the sender agent [7].

Sender:
Formulates a message to send
Obtain start time
For all the messages that have to be sent
Forward message to the receiver
Wait for the response of the receiver
Obtain finish time
The responder role is fulfilled by the Receiver agent:
Receiver:
Wait for the sender's message

One can repeat the measurement for various configurations by putting the agents in different or the same platforms. One can therefore distinguish between the Inter-platform and Intra-platform communications.

Two sub-cases are generated when the whole agents run into a single platform. The first one involves communication among agents who are within the same container. The second one refers to communication among agents who are situated in two different containers [8].

Message Transport Protocols

JADE implements the MTPs (standard Message Transport Protocols) so that interoperability can be promoted among different (with non-JADE) platforms. According to FIPA, the transport protocol definition is contained in every MTP, as well as the message envelope's standard encoding [9].

IV. Results

One of the concerns in the MAS software is finding a way to properly organise and enhance cooperation among the individual agents. Two different approaches were examined. The first one approach is the direct communication, where agent coordination is handled by the agents themselves. The second one refers to assisted coordination, where the agents accomplish coordination by relying on specific system programs. The benefit from direct communication is that this approach does not rely on any other programs existence, capabilities, or biases.

The direct communication popularly has two architectures: contract-net approach and specification sharing.

In the specification sharing approach, the agents give the other agents information about their needs and capabilities so that these agents can then use this information to coordinate their activities.

The second approach is called market-based or contract-net organisation. Thus, agents need to distribute service or transmit requests to other different agents so that they could be used for proposals. The requests are then evaluated and the recipients of the messages submit the bids to the originating agents. The originators use the bids to decide which agents they would task and award contracts to. Often, specification sharing is seen as a more efficient approach compared to the contract-net approach since it lessens the amount of communication that has to be done. However, cost remains as the major dis-advantage of direct communication. As long as there is a small number of an agent, no problems are encountered. However, when there is a larger number, there is an inhibition in the cost of the bids broadcasting and the messages' consequential processing or specifications. As such, in this case, the only alternative is to organise the agents in a way that would prevent such a broadcast. Another drawback has to do with the complexity of implementation. In direct communication schemes, every agent is responsible for negotiating with other agents. Thus, they should have all the codes needed to support such negotiations. The complexity of the application programs may be decreased if the system supplies those capabilities. A common alternative that does not have the two previously mentioned disadvantages is by organising the agents into a federated system. As proposed in the diagram, agents do have direct communication. Instead, they only have communications with system programs called mediators or facilitators.

V. Experiment Results JADE

A. Intra-Platform

- This section will describe the measured round-trip time when the students are found in the same platform. Fig. 5 demonstrates that three results were obtained, which means one for each configuration:
- Abstract must be one paragraph, and no more than 250 words. A minimum of 150 words are suggested, but not mandatory. The abstract must be written as one paragraph, and should not contain displayed mathematical equations or tabular material & should not include references.
- The abstract must be self-contained, without abbreviations, footnotes, or references. It should be a microcosm of the full article.

It can be seen from Fig. 6 that there was low standard deviations during the conduct of the tests. For every graph, the couple's number is found on the X axis, while the related mean round-trip time (avg RTT), presented in the magnitude of milliseconds (ms) is found in the Y axis.

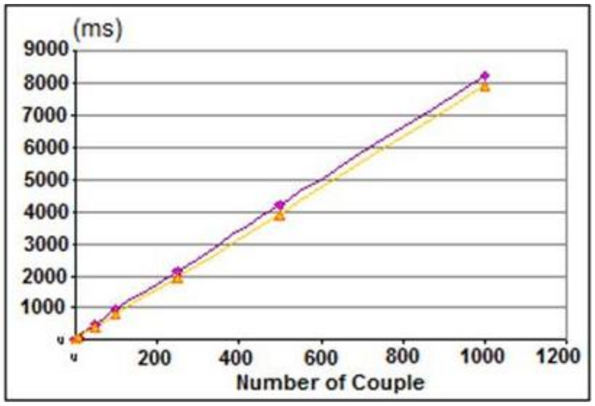


Fig. 5. Round-Trip Time; Intra-Platform Communication.

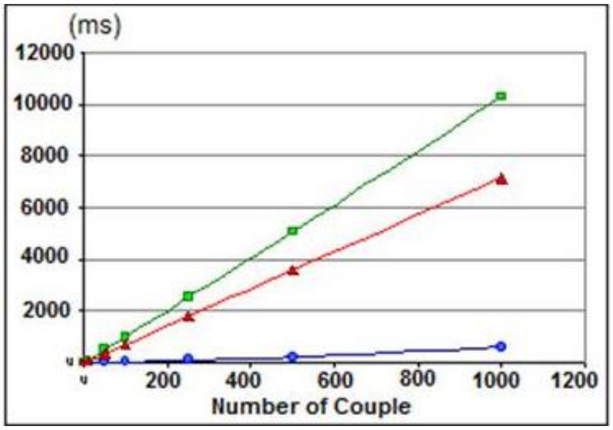


Fig. 6. Round Trip Time; Inter Platform Communication.

At first glance, it was observed that the avg RTT was very low when the students were within the same container. When the students are found inside the same container, JADE optimises the localisation of the students and uses event passing.

Furthermore, for all the cases in the range considered, the middleware demonstrated a linear round-trip time growth as a function of the couple's number.

JADE makes use of MTPs that comply with FIPA specifications. This is an indication that for inter platform communication, JADE needs to use String ACL Encoding to add an envelope to the message before it can be delivered and parsed at the side of the receiver. For instance, the round-trip time for the inter-platform scenario bears great similarity to the intra-platform scenario. Furthermore, the FIPA–MTP that was based on IIOP performed well. Lastly, the FIPA specifications' JADE implementation proved to be very effective.

JADE sends messages using RMI when students on two different containers communicate. It was observed that there was a lower round-trip time for the 2 hosts configuration compared to the one-host case because the load of the computation has to be divided between 2 CPUs. Compared to other cases, the low growth rate of the avg RTT for the same container configuration is an indication that this middleware messaging architecture is capable of supporting the computation high-load and without incurring heavy execution degradation.

B. Inter Platform

Fig. 6 shows the round-trip time among students found on different platforms and where each one is living on a different host.

As a result of the possibility of altering the MTP (Message Transport Protocol), the results that were obtained using the ORB that was made by Sun microsystems went through a comparison with those of ORB acus, which was created by IONA without having any influence on the students' code. According to the results, the performance of ORB acus showed more efficiency.

VI. Discussion

In this system, the agents utilise an agent communication language so that they could document their needs and abilities for their local mediators (facilitators). Furthermore, the agents transmit application level information and requests to their mediators, as well as accept application level requests and information as a response. The mediators use the documentation provided by these agents to convert those application level messages and send them to a place that is suitable or appropriate. Thus, the agents form a federation where they surrender their autonomy to their mediators. The mediators are then responsible for meeting their demands.

VII. Conclusion and Future Work

Further research may require the extension of the solution so that it can be applied to other platforms, such as the mobile platform due to the fact that the JADE system integrates the LEAP library that provides the earth where the operator for cell phone is built and which can be used to deal with genuine servers. Future work may also aim to make this study applicable in answering various contextual analyses.

Agent-based systems have recently gained popularity in industrial and academic fields. AOSE offers diverse conceptual, frameworks, techniques, and notations. Consequently, it provides a platform that offers support for the dynamic, generalisation and autonomous, which in turn helps introduce robustness and the ease of using software methodologies so that challenges can be met and goals achieved. The Menial condition was used to try to control the consequences of c take and use the span postulate as follows: the quest of Coffin-nail configuration was associated to the Ebb method. It is business-like and is therefore dissimilar to multi-agents systems. There are also numerous versions of the methodology, such as ROADMAP, Jade V2, and AUML.

Numerous viewpoints can also be checked and assessed regarding their ability to improve the Jade philosophy, regardless of whether the process is in the execution stage, the stage of framework necessities, within the periods of outline and investigation, or within the Jade system, among others. An answer was obtained regarding the framework necessities that the JADE strategy has disregarded, therefore improving the Jade methodology by giving a formal specification. After a search of the formal specification's part in the system, a way of providing a formal specification of the system via Object Constraints Languages (OCL) was achieved. After selecting the OCL, the unified modelling language models (Class

Diagrams) was selected and combined with the OCL so that it can help introduce a formal system specification into the verification task. After the formal specifications were obtained, these specifications were combined to serve as input during the JADE analysis phase. The combine process was then evaluated via the JADE framework.

This study introduced a correlation with the altercation test from the e-travel principles. It also introduced an antisocial task close to the Etiolate system so that it could help develop agent-based systems in Coffin-nail status through the use of class diagrams that are extendable with OCL constraints. The puppet fit helps enhance the look of the system in terms of rectify performance and management.

REFERENCES

[1] F. D. A. D. Carvalho, A fuzzy clustering algorithm for symbolic interval data based on a single adaptive Euclidean distance. Berlin, Heidelberg: Springer, 10 2006.

[2] L. X. and undefined Lombardo, “Commentary Clustering and General Education at Old Dominion,” 2000. [Online]. Available: http://www.odu. edu/ao/instadv/archive/vol28issue7/clustering.htm.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” ACM computing surveys (CSUR), vol. 31, no. 3, pp. 264–323, 1999.

[4] M. Morandini, D. C. Nguyen, A. Perini, A. Siena, and A. Susi, “Tool-Supported Development with Tropos: The Conference Management Sys-tem Case Study, ondazione Bruno Kessler – IRST Via Sommarive,” Trento, Italy, 2007.

[5] R. Ali, F. Dalpiaz, and P. Giorgini, “A goal-based frame- work for contex-tual requirements modeling and analysis,” Re- quirements Engineering, vol. 15, no. 4, pp. 439–45, 2010.

[6] R. K. Abbas, “Evaluating and Comparing some Agent-Oriented Software Engineering Methodologies,” Journal of College of Education, No1, 2010.

[7] S. Bhardwaj and A. Goyal, ““A comparative analysis of Agent Oriented Requirement Engineering Frameworks”, International journal of computer application,” VOL, vol. 87, no. 8, pp. 11–15, 2014.

[8] E. M and E. R, “Structural analysis of Agent Oriented Methodologies,” International Journal of Information and Computation Technology, VOL, vol. 4, no. 6, pp. 613–618, 2014.

[9] E. S and B. Argente.E, “An Agent-Oriented Software Engineering Methodology to Develop Adaptive Virtual Organizations”, twenty second international joint, 2010.

AUTHOR’S PROFILE

Khalaf Khatatneh Khalaf Khatatneh is an Associate Professor in the School of Computer Science at Al-Balqa Applied University where he has been a faculty member since 2005. He is the Allen School’s Deputy Dean From 2016–2018, he held Khalaf Professorship for Innovation in Computer Science Education.

Khalaf completed his Ph.D. at Rouen University (France) in 2005. His research interests lie in the area of Artificial Intelligence and program-ming languages, ranging from theory to design to implementation. He has collaborated actively with researchers in several other disciplines of computer science, particularly computer architecture on problems at the hardware/software interface.

Khalaf has served on roughly thirty conference and workshop program committees and served as the Program Chair for Koenig 2012. He has served on the Executive Committee, the Steering Committee for the Computer Science Curriculum, and the ACM Education Board.

Khalaf is the instructor for a popular Mupad on undergraduate topics in programming languages and functional programming.

Classifying Cardiotocography Data based on Rough Neural Network

Belal Amin[1]
Department of Technological Development
Tanta Branch
Worker University
Cairo, Egypt

Mona Gamal[2]
Machine Learning and Information Retrieval Department
Faculty of Artificial Intelligence
Kafrelsheikh University
Kafrelsheikh, Egypt

A. A. Salama[3], Khaled Mahfouz[5]
Department of Mathematics and Computer Science
Faculty of Sciences
Port Said University
Port Said, Egypt

I.M. El-Henawy[4]
Computer Science Department,
Faculty of Computers and Information
Zagazig University
Zagazig, Egypt

***Abstract*—Cardiotocography is a medical device that monitors fetal heart rate and the uterine contraction during the period of pregnancy. It is used to diagnose and classify a fetus state by doctors who have challenges of uncertainty in data. The Rough Neural Network is one of the most common data mining techniques to classify medical data, as it is a good solution for the uncertainty challenge. This paper provides a simulation of Rough Neural Network in classifying cardiotocography dataset. The paper measures the accuracy rate and consumed time during the classification process. WEKA tool is used to analyse cardiotocography data with different algorithms (neural network, decision table, bagging, the nearest neighbour, decision stump and least square support vector machine algorithm). The comparison shows that the accuracy rates and time consumption of the proposed model are feasible and efficient.**

Keywords—Accuracy rate; cardiotocography; data mining; rough neural network; WEKA tool

I. Introduction

Dealing with uncertain and inconsistency data in diagnosing diseases is a very challenging problem in medical field. Cardiotcography (CTG) [1, 2, 3 and 4] is one of the most common diagnostic devices in the last few decades representing features of fetus Heart Rate (FHR) and Uterine Contraction (UC) during pregnancy. The features are organized in a dataset with 21 input attributes and 3 classes of fetus state classified into Normal, Suspicious and Pathologic.

CTG is probably the most widely used technique in all obstetrics. It was introduced by Orvan Hess and Ed Hon at Yale University in 1957 [5]. Before that, the only device used was a stethoscope to determine fetal status and maternal uterine contractions. Therefore, the birth process was as a black box. Before 2008, fetal heart rate was classified as either reassuring or non-reassuring. The NICHD Workgroup [6] proposed a terminology of a three-tiered system to replace the older; they were normal, indeterminate and abnormal. In 2015 FIGO [6] updated the terminology of CTG monitoring device into normal, pathological and suspicious states. The device has several benefits for patients [7, 6, 5]. For example, it helps doctors monitoring more than one patient at the same time, predicting whether the mother needs a cesarean section or not, detecting low and high risk for patients in labour to make decisions quickly. Hence, Fetal Heart Rate (FHR) monitoring remains a widely used method for detecting changes in fetal oxygenation that can occur during labor.

Data mining provides various classification techniques with a suitable accuracy rate and time for such medical data to make decisions or discover patterns in datasets.

In the last decades, the researcher provided many papers on data mining techniques supporting bioinformatics to classify and process data with efficient performance. Dr C Sunder [1] used supervised artificial neural network and support vector machine to classify the CTG dataset depending on training data. But the model didn’t have a good performance to classify a suspicious state as the other two states normal and pathological. Dr Ahmed Abou El-Fetouh [8] used hybrid rough neural network model to analyze the performance of breast cancer classification using different sizes of training data. The paper used WEKA [9, 10] tool to measure accuracy rate of Neural Networks and compare the results. But it didn’t estimate the consumption time of the RNN [8, 11] model and didn’t use more algorithms to compare with the proposed model. Dr Suman [3] used WEKA [9, 10] mining tool to analyze classification techniques like (neural network, Bayesian classification and decision tree) to provide which technique has the best and efficient performance. Comparison among different algorithms determines that each algorithm performs the best result according to its parameters, but he did not determine which one was the best in general to use as classification technique [9], Dr Divya Bhatnagar [10]. Provided analyses of CTG data set and generated classification rules to identify normal, suspicious and pathological cases using WEKA classifiers. He didn’t apply simulation for his results or provide hybrid models for improving classification accuracy rate. Z. Cömert [12] presented the comparative metrics of five machine learning techniques such as Artificial Neural Network (ANN) [1, 2, and 13], support vector machine [14], extreme learning machine [15], radial basis function network [16] and random forest [17]. He found that ANN technique is the most efficient in the sensitivity and specificity measures. Dr Mona Gamal[18] used hybrid model of fuzzy rough feature selection and rough neural networks to classify dataset of breast cancer and measure accuracy rate and consumed time of processing data.

The importance of the proposed model is to present a solution of limitations in the previous researches. Providing a good performance to classify all states of fetus heart rate. Also, comparing its results with various algorithms to prove that it satisfies a good accuracy rate in suitable time consumption. And providing analysis of CTG attributes using a WEKA application to visualize it.

The proposed model depends on Rough Neural Network (RNN) [8, 11] which is built on a neural network structure [1, 2 and 13] and rough sets theory [8, 11]. RNN is characterized by various advantages such as the ability to deal with fault tolerance, simplicity and relief of structure, parallel processing of dataset and self-adapted. In addition, RNN advantages of rough set in performing sustainable amount of uncertain data and reduction attributes without losing information. RNN is composed of multilayers input, hidden and output layers. The simulated model measures accuracy rate and time consumption on (CTG) dataset.

The paper is organized as follows; Section 2 presents information about CTG device and its role in diagnosing fetal status. Section 3 provides the proposed model, algorithm and its benefits. Section 4 presents an experimental result of the model and analysis of other classification techniques in comparison. At the end, conclusion and future work are documented in section 5.

II. Cardiotocography

Cardiotocography [2] is common medical devices; many re-searches analyze datasets to achieve improved accuracy in diagnosing the state of fetal heart rate under uncertain situations. The device produces a simultaneous recording and traces patterns of the FHR and the UC during pregnancy period and before delivery. Now the Cardiotocography readings are organized and stored for medical researches.

The CTG dataset consists of measurements of FHR and UC for the fetus, the important features of Cardiotocograms classified by an obstetricians' expert, and the data set is available publicly at the data mining repository of University of California Irvine (UCI) [4]. (Last accessed April 2019). Data set was split into training data and testing data with percentages 70% and 30% respectively.

The data set has 21 attributes and classified according to the FHR pattern or fetal state class code [3, 4]. In this study, fetal state class code is used as the target attribute instead of FHR pattern class code and classification into one of three groups Normal, Suspicious or Pathological (NSP) classes. The dataset includes a total of 2126 samples. Attributes description is given in Table I.

TABLE I. CTG Data Set Attributes Description

CTG data set attributes description	
Attribute	**Description**
LB	Fetal Heart Rate baseline (beats per minute)
AC	number of accelerations per second
FM	number of fetal movements per second
UC	number of uterine contractions per second
DL	number of light decelerations per second
DS	number of severe decelerations per second
DP	number of prolonged decelerations per second
ASTV	percentage of time with abnormal short term variability
MSTV	mean value of short term variability
ALTV	percentage of time with abnormal long term variability
MLTV	mean value of long term variability
Width	width of FHR histogram
Min	minimum of FHR histogram
Max	maximum of FHR histogram
Nmax	number of histogram peaks
Nzeros	number of histogram zeros
Mode	histogram mode
Mean	histogram mean
Median	histogram median
Variance	histogram variance
Tendency	histogram tendency
CLASS	FHR pattern class code (1 to 10)
NSP	fetal state class code (N = normal; S = suspicious ; P = pathologic)

III. Proposed Model

The proposed model uses RNN [8, 11], which depends on combining Neural Network (NN) [1, 2, and 13] and rough set theory [8, 11]. The proposed model applies the supervised learning model of the RNN and formed from three consecutive phases which are preprocessing, training and testing phases as in the following:

1) *Preprocessing phase:* where medical dataset is normalized to avoid anomaly values of features and improve the efficiency of medical data in implementation stage.

2) *Training phase:* where the RNN is trained to reach best weights helps in discovering patterns of data and reduce absolute error by using a feed forward algorithm, and back propagation algorithm to update upper and lower weights to reach a better classification of CTG data set.

3) *Testing phase:* where the trained RNN is measured against new instances of data to calculate the accuracy rates using the relation: Accuracy Rate = 1 – absolute error. Moreover, the time consumption is determined to prove the performance of the proposed model.

The RNN structure replaces the traditional neuron by two neurons (lower neuron, upper neuron) to represent lower and upper approximations of each attribute in the CTG data set, its structure formed from 4 layers input, 2 hidden and output layers. The hidden layers have rough neurons, which overlap and exchange information between each other, While the input and output layers consists of traditional neurons as in Fig. 1.

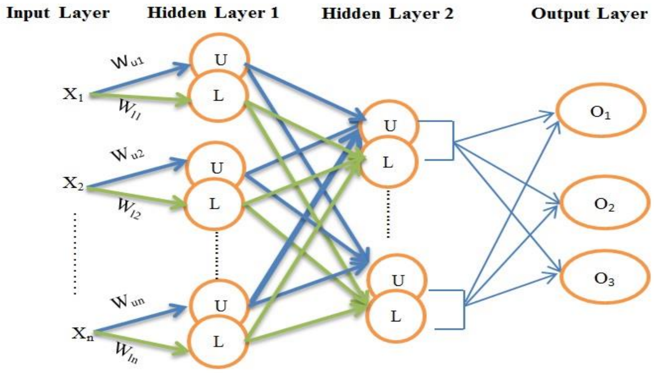


Fig. 1. Rough Neural Network (RNN) Structure.

Input layer is composed of neuron for each data attribute. The output layer represents the three FHR classes, the hidden layers rough neurons are determined by the Baum-Haussler rule [19].

$$N_{hn} = \frac{N_{ts} * Te}{N_i + M_o} \quad (1)$$

Where N_{hn} is the number of hidden neurons, N_{ts} is the number of training samples, Te is the tolerance error, N_I is the number of inputs (attributes or features), and N_O is the number of the output.

During training process, the normalized input data is multiplied by its weight and computed in sigmoid activation function.

$$f(x) = \frac{1}{1+e^{-\lambda x}} \quad (2)$$

Steps of the proposed model architecture:

Step I: preprocessing phase

1. Read features of each objects in dataset
2. Normalize all values of data by equation

$$Nor = \frac{X - \min}{\max - \min} \quad (3)$$

Step II: Training phase

3. Initialize random (upper, lower) weights of network
4. Feed forward of attribute values and multiply in both direction (U_w, L_w)
5. Compute (I_U,I_L) of hidden layers by relations:

$$I_{Ln} = \sum_{j=1}^{n} W_{Lnj} O_{nj} \quad (4)$$

$$I_{Un} = \sum_{j=1}^{n} W_{Unj} O_{nj} \quad (5)$$

6. Compute (O_U , O_L) of hidden layers by relations:

$$O_{Ln} = \text{Min} (f (I_{Ln}), f(I_{Un})) \quad (6)$$

$$O_{Un} = \text{Max} (f (I_{Ln}), f(I_{Un})) \quad (7)$$

7. Check fetus according to comparing between actual output (T) and output value (O), where output represent by

$$O = O_{Ln} + O_{Un} \quad (8)$$

8. If output is error, then use back propagation algorithm, and compute error.

$$\Delta = T - O \quad (8.1)$$

9. Update (upper, lower) weights of network by derivation of activation function: new weight = old weight + (Δ * η *derivative* activation of(input)) (10) where η is learning rate of model
10. Repeat 5, 6, 7, 8 and 8.1 until reduction error as possible as.

Step III: Testing phase

Classify new sample of objects and determine the accuracy rate of the model by using relation Accuracy = 1–absolute error, also calculate time consumption in model processing.

The flowchart of the hybrid proposed model is shown in the following Fig. 2.

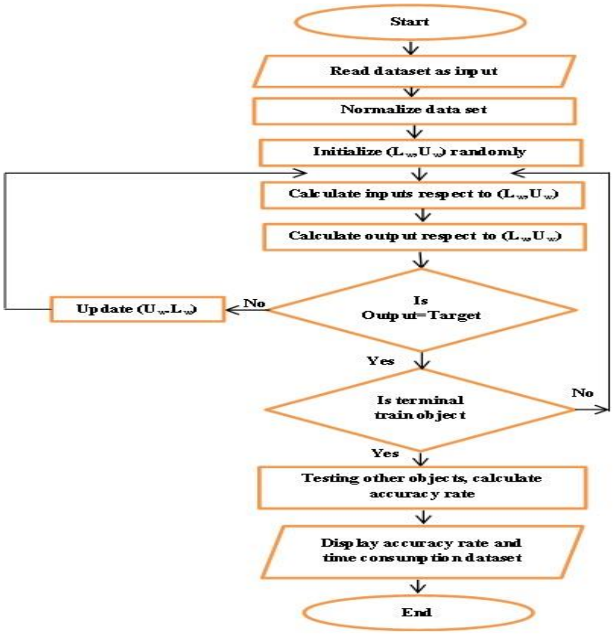

Start
Read dataset as input
Normalize data set
Initialize (L_w,U_w) randomly
Calculate inputs respect to (L_w,U_w)
Calculate output respect to (L_w,U_w)
Is Output=Target
No
Update (U_w,L_w)
Yes
Is terminal train object
No
Yes
Testing other objects, calculate accuracy rate
Display accuracy rate and time consumption dataset
End

Fig. 2. Flowchart of Rough Neural Network (RNN) Algorithm Steps.

IV. Experimental and Results

CTG is an important medical device that has 21 features to determine the state of fetus heart rate and uterine contraction at the same time. Obstetricians could classify the state of fetal as normal, pathologic or suspicious state according to values of its features. So it’s vital to visualize [20, 21] of cardiotcography device features by using WEKA version 3.7.2 [9, 10] application as in Fig. 3. The attribute is drawn to illustrate a visual qualitative understanding of the distribution.

A boxplot is a statistical way to summarize large amounts of data of each feature and display each of minimum, maximum, range, median and distribution of data. Also, it shows the symmetry of data, the upper and lower quartiles, which represent the numbers above and below the high and lower quarters of the data. The CTG data set boxplot is presented in Fig. 4.

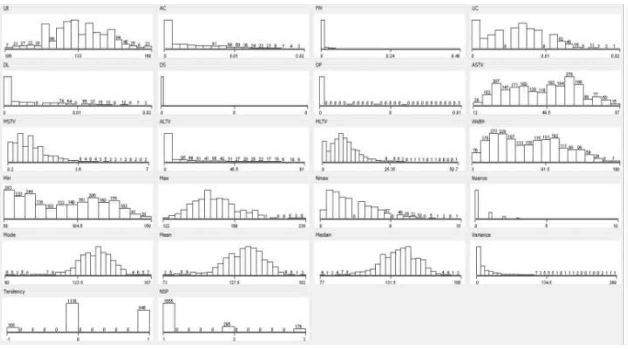


Fig. 3. Distribution of CTG Features.

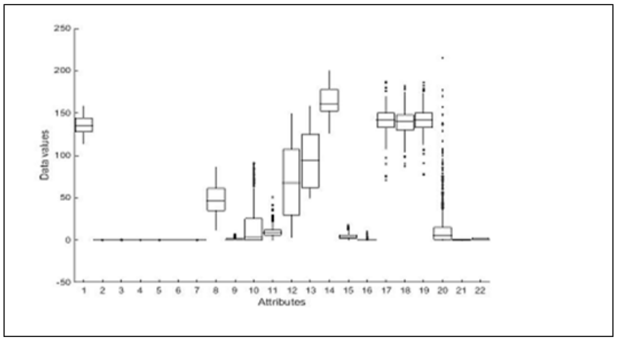


Fig. 4. Boxplot of CTG Features.

The proposed model is implemented in the structure of the rough neural network (RNN) as it's one of the best models in processing uncertain medical data. It is composed of four layers input, two hidden and output layers. The input layer is formed from conventional neuron, while hidden layers are formed from pairs of neurons called upper and lower neurons, which overlap to exchange information and using the interval of values. The output layers formed of conventional neurons represent classes of CTG data set which could be normal, suspicious and pathological states. The proposed model is implemented by C# language. In Windows 7 by specification device, processor Intel ®core™ i5, Ram 4 GB, 64- bit operating. The processing is in three steps which are the preprocessing phase that normalizes features values to avoid anomalies, the training phase, which achieves the RNN learning by using back propagation algorithm to update the weights on networks. The third phase is testing which measures the accuracy rate of the classifier and time consumption during processing the model, The CTG data set is divided into training and testing datasets with a percentage of 70% as training data to learn machine and 30% as testing data to compute accuracy rate of the model.

The WEKA [9,10] tool to analysis CTG dataset using different data mining classifiers such as Nearest Neighbor [22,23], Neural Network[1,2,13], Bagging [24], Decision Table [25,26,27], Decision Stump [28,29] and Least Square Support Vector Machine algorithm [3,30] and compute accuracy rate as in the following Table II.

As shown in the table our model achieves the best accuracy rate compared to other classifiers and it satisfies more efficiency performance. Fig. 5 represents a chart of them as the following.

The time consumption in second of the model is computed and compared to other classifiers as in the following Table III and Fig. 6 observed the RNN model has an acceptable consumption time.

TABLE II. COMPARISON BETWEEN DIFFERENT DATA MINING ALGORITHMS IN ACCURACY RATE

Algorithm	Accuracy Rate
Rough Neural Networks	92.95 %
Nearest Neighbor	84.99 %
Neural Network	83.12 %
Bagging	85.15 %
Decision Table	77.85 %
Decision Stump	66.05 %
Support Vector Machine for regression	74.92 %

TABLE III. COMPARISON BETWEEN DIFFERENT DATA MINING ALGORITHMS IN TIME CONSUMPTION

Algorithm	Accuracy Rate
Rough Neural Networks	14.25
Nearest Neighbor	0.03
Neural Network	10.01
Bagging	0.39
Decision Table	0.35
Decision Stump	0.04
Support Vector Machine for regression	18.37

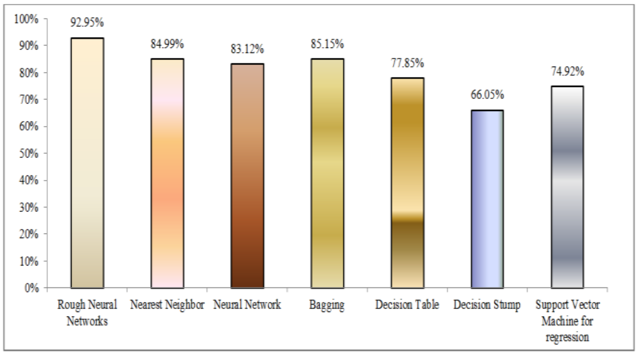


Fig. 5. Comparison between different Data mining Algorithms in Accuracy Rate.

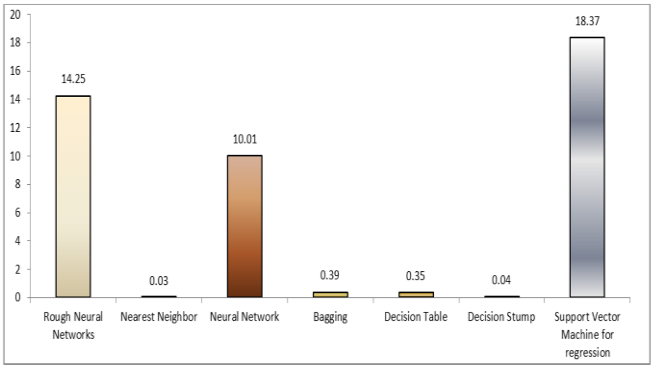


Fig. 6. Comparison between different Data mining Algorithms in Time Consumption.

V. Conclusion and Future Work

Features of fetus Heart Rate (FHR) and Uterine Contraction (UC) during pregnancy are very important in monitoring fetus and mother's health. Data mining provides important techniques for dealing with uncertain medical data. Rough Neural Network classifier is based on neural network and rough set theory. RNN is a well-tested algorithm which satisfies efficiency and provides a good diagnosing of diseases rapidly with high accuracy. RNN structure is composed of rough neurons that manage the upper and lower boundaries in the input and hidden layer instead of traditional neuron with full connection between upper neurons, and lower neurons. During training, RNN learns its weights basing on the back propagation algorithm to updates upper and lower weighs boundaries of input and hidden layers. Through the testing phase, the system measures the accuracy of data and time consumption of the processing model.

The paper presents a distributed and boxplot visualization of CTG features by WEKA tool. Also, the paper provides an implementation of the proposed model, computes the accuracy rate of CTG data set based on absolute error and time consumption. Comparisons between the proposed model and different data mining algorithms such as Nearest Neighbor, Neural Network, Bagging, Decision Table, Decision Stump and Least Square Support Vector Machine algorithm proved the feasibility of the RNN in classifying the CTG data basing on accuracy rate in suitable time.

The future work, several improvements should be made on the accuracy rate of the proposed model technique and apply other data mining techniques. Also, feature selection algorithms would be applied to remove irrelevant features.

References

[1] C. Sunder, M. Chitradevi, and G. Geetharamani, "Incapable of identifying suspicious records in CTG data using ANN based machine learning techniques", Journal of Scientific& Industrial Research,vol 73, , Augest 2014, pp. 510-516.

[2] C. Sunder, M. Chitradevi, and G. Geetharamani," Classification of Cardiotocogram Data using Neural Network based Machine Learning Technique", International Journal of Computer Applications, vol 47, June 2012.

[3] Jagannathan D., "Cardiotocography, "A Comparative Study between Support Vector Machine and Decision Tree Algorithms", International Journal of Trend in Research and Development, vol 4, February 2017.

[4] http://archive.ics.uci.edu/ml/datasets/cardiotocography.(Accessed February 2019).

[5] https://www.sciencedirect.com/topics/medicine-and-entistry/cardiotocography (Accessed February 2019).

[6] https://en.wikipedia.org/wiki/Cardiotocography.(accessed February 2019).

[7] R. J. Pardeshi, D. V. Gopalghare, M. Sase, and P. P. Panigrahi, "Carditocograghy In Early Labour- A Screning Test For Prediction Of Fetal Outcome", International Journal of Innovative Medicine and Health Science, vol 4, pp. 81-85,2015.

[8] A. A. El-Fetouh, M. Gamal, "An Intelligent Model in Bioinformatics based on Rough-Neural Computing", International Journal of Computer Applications, vol 64, Febuary 2013.

[9] Suman, and P. Mittal, "A Comparative Performance Analysis of Classification Algorithms Using Weka Tool Of Data Mining Techniques", International Journal of Computer Science and Information Technologies, Vol. 5, 2014.

[10] D. Bhatnagar., P. Maheshwari., "Classification of Cardiotocography Data with WEKA", International Journal of Computer Science and Network, vol 5, April 2016.

[11] S. Ding, J.Chen, X Xu, and J. Li," Rough Neural Networks: A Review", Journal of Computational Information Systems,vol 7, 2011.

[12] Z. Cömert, and A.F. Kocamaz, "Comparison of Machine Learning Techniques for Fetal Heart Rate Classification", 3rd International Conference on Computational and Experimental Science and Engineering (ICCESEN 2016), vol 132, 2017.

[13] David K, "A Brief Introduction to Neural Network", http://www.dkriesel.com/en/science/neural_networks, 2005.

[14] W. S. Noble, "What is a support vector machine?", Nature Biotechnology, vol 24 , Dec 2006.

[15] S. Ding, X. Xu, and R. Nie, "Extreme learning machine and its Applications", Journal of Neural Computing and Applications, Vol 25, Pages 549-556,Sep 2014.

[16] Q. Que, and M. Belkin," Back to the Future: Radial Basis Function Networks Revisited", 19th International Conference on Artificial Intelligence and Statistics (AISTATS), vol 51, 2016.

[17] G. Biau, "Analysis of a Random Forests Model", Journal of Machine Learning Research, vol 13, pp 1063-1095, 2012.

[18] M. Gamal, "Medical Diagnostic System Basing Fuzzy Rough Neural-Computing for Breast Cancer", International Conference on Advanced Intelligent Systems and Informatics, Springer, pp. 467-479, 2016.

[19] E.Baum, D.Huassler," what size net gives valid generalization", the Neural Information Processing Systems Conference., 1988.

[20] R. Dawson," How Significant Is A Boxplot Outlier?", Journal of Statistics Education, vol 19, 2011.

[21] H. Wickham and L. Stryjewski, "40 years of boxplots", Technical report, http://vita.had.co.nz/papers/boxplots.pdf, 2011.

[22] Y. N. Law, and C. Zaniolo., "An Adaptive Nearest Neighbor Classification Algorithm for Data Streams", Springer-Verlag Berlin Heidelberg, pp. 108-120, 2005.

[23] S. B. Imandoust, and M. Bolandraftar," Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background", International .Journal of Engineering Research and Applications, vol 3, Sep-Oct 2013, pp 605-610.

[24] P. Shrivastava, and M. Shukla," Uses the Bagging Algorithm of Classification Method with WEKA Tool for Prediction Technique", International Journal of Advanced Computational Engineering and Networking, vol 2, December 2014.

[25] H. Lu, and H. Liu," Decision Tables: Scalable Classification Exploring RDBMS Capabilities", Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000.

[26] J. Kaur, and Seema Baghla, "Modified Decision Table Classifier by Using Decision Support and Confidence in Online Shopping Dataset", International Journal of Computer Engineering & Technology (IJCET), vol 8, 2017, pp. 83–88.

[27] P. Sewaiwar, and K. K. Verma., "Comparative Study of Various Decision Tree Classification Algorithm Using WEKA", International Journal of Emerging Research in Management &Technology, Oct 2015.

[28] Ayinde A.Q, Adetunji A.B, Bello M and Odeniyi O.A., "Performance Evaluation of Naive Bayes and Decision Stump Algorithms in Mining Students' Educational Data", IJCSI International Journal of Computer Science Issues, vol 10, July 2013.

[29] T. Kannapiran, and M. Wadhawa," Analysis and Comparison Study of Data Mining Algorithms Using RAPID Miner", International Journal of Computer Science, Engineering and Applications (IJCSEA), vol 6, Feb 2016.

[30] E. YJlmaz, and Ç. KJlJkçJer," Determination of Fetal State from Cardiotocogram Using LS-SVM with Particle Swarm Optimization and Binary Decision Tree", Computational and Mathematical Methods in Medicine, 2013.

Hospital Queue Control System using Quick Response Code (QR Code) as Verification of Patient’s Arrival

Ridho Hendra Yoga Perdana[1], Hudiono[2], Mochammad Taufik[3], Amalia Eka Rakhmania[4]
Rohman Muhamad Akbar[5], Zainul Arifin[6]
Electronic Engineering Department
State Polytechnic of Malang, Malang, East Java, Indonesia

***Abstract*—Hospital is an organization that primarily provides services in the form of examination, treatment, medical treatment and other diagnostic measures required by each patient in the limits of the technology and the means provided by the hospital. So that patients get the maximum service, hospitals have to provide the best services, one of the services that are being highlighted is the queue system. This study proposed a queue control system at the hospital using a quick response code (QR Code) as verification of the patient's arrival that serves to speed up the administrative process and provide effective services, to facilitate the patient in terms of the queuing line (registration) at the hospital. The queue control system uses the website as a registration and database repository for patient data. This website has two kinds of display form, as a patient or a doctor. This kind of display is useful to speed up the administrative process, so that registration can be done at any time by the patient, with patient entering/filling data themselves and their complaints or medical records that are automatically available on the website, so that patients do not need to come to the hospital with a medical record. Furthermore, the website will filter the data entered by the patient as the patient's complaints, so that patients can choose the hospital that provides services for patients’ illness complaints which are also showing hospital addresses to facilitate patients in finding the location of the hospital. After the administration process on the website, the system will provide a QR Code as a form of verification for patients who have registered, with their data already exists in the database. When the patient arrives at the hospital, the patient only needs to scan the QR code that already exists, without taking care of the administration back because at the time the system gives a QR Code, indicating that patients already had a hospital destination along with the queue number. It is expected that the queue control system implemented at the hospital using a quick response code (QR Code) as verification of the patient's arrival, could speed up the administrative process to be more effective and efficient.**

***Keywords*—Queue; hospital; patient; quick response code; registration**

I. Introduction

In Indonesia, there is a service that is often used by people to support/serve complaints about public health. This kind of service/support has a primary goal of providing services in the form of examination, treatment, medical treatment, and other diagnostic measures required by each patient in the limits of technology and infrastructure provided by the hospital. The hospital also has rules in serving patients, such as a room or a bed of patients [1][2], the quality of drugs, and the facilities provided [3]. So that patients get the maximum service, hospitals have to provide the best services, one of the services that are being highlighted is the queue system because patients often complain about the length of the queue to the checkout process [4] [5]. In order for hospital operations can run effectively and efficiently, it needs development in the infrastructure field [6]. One is through the implementation of technology.

The development of technology is currently more advanced and has a great advantage, so we need a system that serves to support the advancement of technology that could produce the required information faster and more convenient than the manual method [7]. One example is the Quick Response Code (QR Code), which is a type of code that is able to convey information quickly and get a quick response anyway. Unlike the bar code, which only stores information horizontally, a QR code is able to save the information horizontally and vertically, and therefore automatically a QR code can hold more information than a bar code [8][9]. QR codes have high capacity in the data encoding, which is capable of storing all kinds of data, besides the QR code has a smaller screen than the barcode. This is because the QR code can hold the data horizontally and vertically, so automatically the size of the display of the QR code image can be only 10% of the size of a barcode. Not only that QR codes are also resistant to damage, because the QR code is able to correct the error up to 30%. Although some symbol of the QR code is dirty or damaged, the data can still be stored and read [10][11]. Three square-shaped marks in three corners have a function so that the symbol can be read with the same results from any angle throughout 360 degrees [12].

This study proposed a queue control system at the hospital using a quick response code (QR Code) as verification of the patient's arrival. The system provides an effective service to facilitate the patient in terms of administration (queue). The system uses the website as an interface that can be accessed by the user anytime and anywhere. During the registration process line, the patient will enter the data themselves and their complaints on a website, and then they will get their medical records that are automatically stored in the patient account, so the patient does not need to come to the hospital with a medical record. Then the system will filter the data that has been entered by the patient, such as personal data, patient’s disease complaints, hospitals that provide services for such diseases, as well as hospitals desired address, in order to facilitate the patient in finding the location of the hospital [13]. After the administration process on the website, patient data will be stored in the database and the system will provide a QR Code as verification for patients who have registered. Data stored on the QR Code including hospital lab data that are in accordance with the patient's medical record, which has been filtered by the system along with the queue number, will be automatically processed by the system and generate data stored in the patient's QR Code. When arriving at the hospital, the patient only needs to scan the QR code that has been downloaded, without doing the re-administration.

II. Methodology

A. Phase 1; Requirements Planning

At the stage of making a line control system with Quick Response Code in this hospital is divided into several sections, this section includes several parts:

1) Hospital design: This section describes the workflow tool used in hospitals. It also explains how the doctor is doing the examination process and to provide action on the patient.

2) User: This section describes the concept created for patient registration, booking, arrival verification, as well as the examination of patients by doctors.

3) Relation database: This database describes the concept of the queue control system with Quick Response Code, this design also includes web design and database design.

B. Phase 2; Hospital Design

Hospital's main purpose for the patient is to obtain medical care, while the hospital workflow process can be seen in Fig. 1.

Fig. 1 is a workflow in hospitals, the first step for the patient is to scan the QR Code to verify the arrival of the patient in order to indicate that the patient has arrived at the hospital. The patients QR code will be scanned with QR Code sensors that are available at the front desk. Further information will appear on the LCD, this information will appear when the patient has already verified for their arrival at the hospital, the information on the LCD is useful to simplify the patient action to get the examination at the hospital. The LCD contains information such as the name of the lab, patient id, queue number, and patient's name. Then the examination process will be done by a doctor, where the doctor will examine the patient while providing certain measures to the patient.

1) System design: This design system is a general description of how the tool works from the queue control system and for the flow can be seen in Fig. 2.

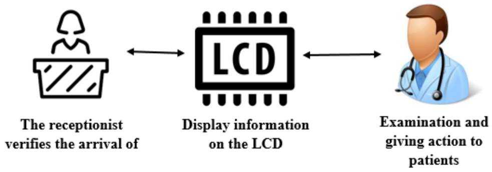

The receptionist verifies the arrival of
LCD
Display information on the LCD
Examination and giving action to patients

Fig. 1. Work Process Hospital.

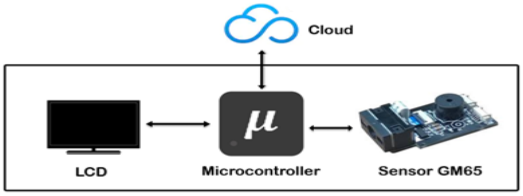


Fig. 2. Machine Operating System.

Fig. 2 shows the design of the tool, how this tool works. First, the GM65 sensor will scan the patient's QR Code to verify the patient's arrival at the hospital, the process of scanning this data is by matching the code that has been obtained by the patient with the code in the database. Second, the microcontroller works to control the process of starting the QR Code scanner, verifying the patient's data, retrieving data in the database, and displaying data on the LCD screen. Third, this cloud function is to connect hardware with software (website) so that the patient data can be stored in the database. Fourth, the LCD will display information to make it easier for patients to go to a certain lab in the hospital.

2) Hardware design: This hardware work process can be seen in Fig. 3.

Fig. 3 descriptions are:

a) Quick Response Code (QR Code): Patients Quick Response Code (QR Code) is used to verify the arrival of patients. The patients can only get a QR Code after registering on the website; this code will later be scanned to verify the arrival of patients at the hospital.

b) Scan Quick Response Code (QR Code) in hospital: To verify the arrival of patients at the hospital the patient must scan the Quick Response Code (QR Code) at the hospital. This process of scanning data is by matching the code obtained by the patient with the code in the database. If the patient does not register on the website, the patient will not get the Quick Response Code (QR Code). If the code matches, the process will proceed to Arduino Uno.

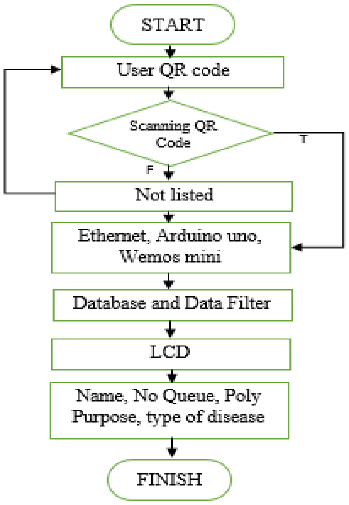

START
User QR code
Scanning QR Code
T
F
Not listed
Ethernet, Arduino uno, Wemos mini
Database and Data Filter
LCD
Name, No Queue, Poly Purpose, type of disease
FINISH

Fig. 3. Tool Design (Hardware).

c) Ethernet and Arduino Uno: Ethernet is used to connect mobile devices. Which then the mobile device will scan user/patient Quick Response Code (QR Code). Arduino itself is used to match Quick Response Code (QR Code) and filter patient registration data from the database. After filtering, the results of the data selection will be displayed on the Liquid Crystal Display (LCD) screen.

d) Database: This database is used to store patient medical records and patient registration data. This database will be accessible to patients and hospitals. The hospital will access this database to retrieve patient data in accordance with Quick Response Code (QR Code) which is owned by the patient during verification arrival. This database is in the form of a place hosting/server.

e) Wemos Mini: Wemos Mini hardware is used to connect devices to the Internet so that it can access the database that has been provided.

f) Data Filter: This data filter is done by Arduino Uno, because the data filter includes location, disease category, and hospital lab destination. Results of this data filter will be displayed on the Liquid Crystal Display (LCD) screen to facilitate patients in terms of the queue process.

g) Liquid Crystal Display (LCD): Liquid Crystal Display (LCD) is used to display data from the Arduino filter results. The data displayed is in the form of the patient's name, hospital lab destination, and queue number.

Fig. 4 shows the scanning process of the QR Code with the GM65 sensor.

This Quick Response Code is used to verify the arrival of patients in the hospital, namely with GM 65 sensors as shown in Fig. 4.

3) Doctor design: The workflow process of doctors at the hospital can be seen in Fig. 5.

Fig. 5 descriptions are:

a) Login: Doctors must be logged on to the website first before starting the examination of the patient. The login process can be done by entering the username and password that is owned by the doctor.

b) Scan the Patient QR Code: QR Code Scan is done so that doctors can enter the patient's medical record in question and give action to these patients. In giving action, doctors can go through the website or application on Android.

c) Examination and giving action to the patient: The examination will be carried out when the doctor has entered the patient's medical record and gives action to the patient. After that, the doctor will fill in the patient's medical record that has been available on the action form.

Fig. 6 shows the process flow from the doctor. The doctor will log in first, then the doctor will go into the patient's history to provide follow-up after the examination. Then, the doctor can see the patient's history that is available on the website's features.

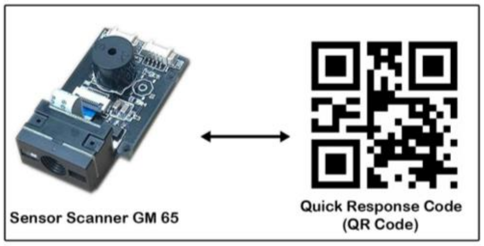


Fig. 4. Doctor Activity Diagram.

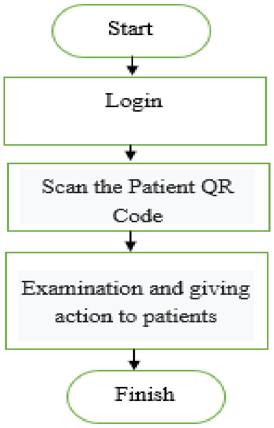


Fig. 5. Doctor Workflow Process.

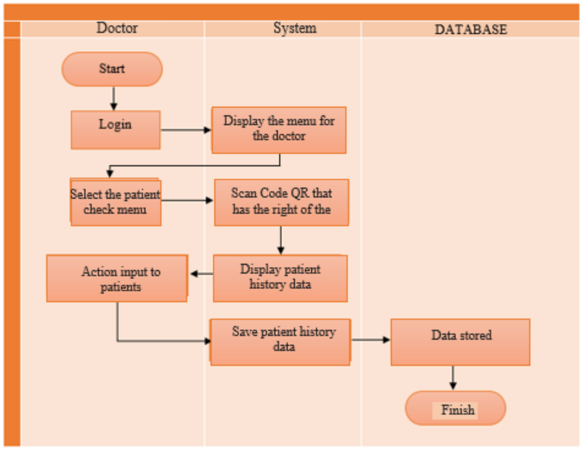

Doctor
System
DATABASE
Start
Login
Display the menu for the doctor
Select the patient check menu
Scan Code QR that has the right of the
Action input to patients
Display patient history data
Save patient history data
Data stored
Finish

Fig. 6. Doctor Activity Diagram.

C. Phase 3; user Design

The input of the queue control system and the patient's workflow process is described in Fig. 7.

Fig. 7 shows a flowchart about the block diagram of the queue control system. The first step is creating an account/login as a patient. Ordering a queue at the hospital can be done by accessing the website and registering a new account to log in to the website to place an order. Registration form consists of name, address, date of birth, username, e-mail, and password which will be filled by the patient to create a new account, then the patient will enter the website by entering the username and password that has been registered.

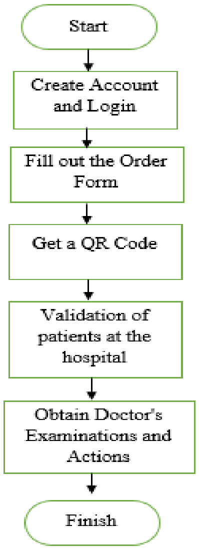

Start
Create Account and Login
Fill out the Order Form
Get a QR Code
Validation of patients at the hospital
Obtain Doctor's Examinations and Actions
Finish

Fig. 7. Queue Control System Block Diagram.

The second step is to fill in the order form. The patient registration form is on the order menu, this order form is filled in by the patient to order queues at the hospital. In the order form, there are fillings for order date, city, hospital, hospital lab, and complaints that will be filled by the patient. After filling out the order form, patients can choose the doctor's schedule that has been provided by the hospital. After pressing order, the patient will get a QR Code. This QR Code will be used to verify arrival at the hospital. This Quick Response Code (QR Code) will be obtained by the patient when the patient has placed an order on the website, the QR code on the website can be downloaded by the patient in a PDF format.

The third step is to verify the arrival at the hospital. To verify the arrival of the patient in the hospital, the patient must scan his QR code (which was obtained during online registration), the QR Code sensor scanner at the hospital. Then on the LCD screen will display hospital lab name, patient id, queue number, and patient name to make it easier for patients to carry out examinations at the hospital.

The fourth step is to get a doctor's examination and action. After the patient verifies the arrival, the patient will undergo an examination and get action from the doctor, and the doctor will scan the patient's QR code to enter the patient's medical record data. Patients can see the examination history of the patient's medical record menu on the website.

Fig. 8 shows the process flow of the user/patient. Patients will register (not already have an account) as well as ordering the queue at the hospital designated by using the website. Patients will enter the data themselves in the provided form and will get a QR code for arrival verification.

D. Phase 4; Relation Database Design

This database is used for data storage on the server. And this database is also used as one of the supporting systems for the queue system at the hospital, and the database workflow can be seen in Fig. 9.

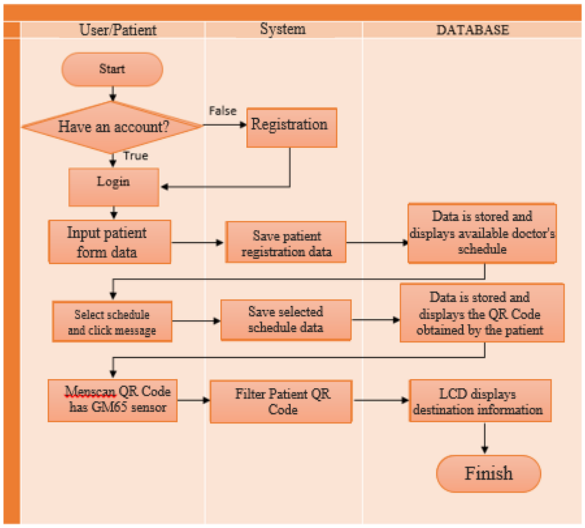
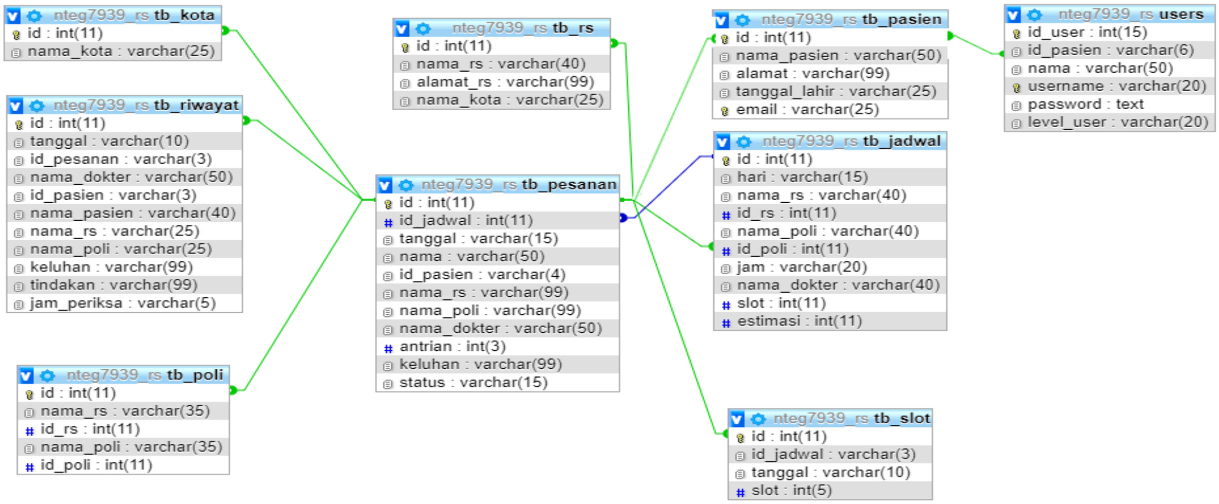

User/Patient
System
DATABASE
Start
Have an account?
False
Registration
True
Login
Input patient form data
Save patient registration data
Data is stored and displays available doctor's schedule
Select schedule and click message
Save selected schedule data
Data is stored and displays the QR Code obtained by the patient
Menscan QR Code has GM65 sensor
Filter Patient QR Code
LCD displays destination information
Finish

Fig. 8. User Activity Diagram.


nteg7939_rs tb_kota
id : int(11)
nama_kota : varchar(25)

nteg7939_rs tb_riwayat
id : int(11)
tanggal : varchar(10)
id_pesanan : varchar(3)
nama_dokter : varchar(50)
id_pasien : varchar(3)
nama_pasien : varchar(40)
nama_rs : varchar(25)
nama_poli : varchar(25)
keluhan : varchar(99)
tindakan : varchar(99)
jam_periksa : varchar(5)

nteg7939_rs tb_poli
id : int(11)
nama_rs : varchar(35)
id_rs : int(11)
nama_poli : varchar(35)
id_poli : int(11)

nteg7939_rs tb_rs
id : int(11)
nama_rs : varchar(40)
alamat_rs : varchar(99)
nama_kota : varchar(25)

nteg7939_rs tb_pesanan
id : int(11)
id_jadwal : int(11)
tanggal : varchar(15)
nama : varchar(50)
id_pasien : varchar(4)
nama_rs : varchar(99)
nama_poli : varchar(99)
nama_dokter : varchar(50)
antrian : int(3)
keluhan : varchar(99)
status : varchar(15)

nteg7939_rs tb_pasien
id : int(11)
nama_pasien : varchar(50)
alamat : varchar(99)
tanggal_lahir : varchar(25)
email : varchar(25)

nteg7939_rs users
id_user : int(15)
id_pasien : varchar(6)
nama : varchar(50)
username : varchar(20)
password : text
level_user : varchar(20)

nteg7939_rs tb_jadwal
id : int(11)
hari : varchar(15)
nama_rs : varchar(40)
id_rs : int(11)
nama_poli : varchar(40)
id_poli : int(11)
jam : varchar(20)
nama_dokter : varchar(40)
slot : int(11)
estimasi : int(11)

nteg7939_rs tb_slot
id : int(11)
id_jadwal : varchar(3)
tanggal : varchar(10)
slot : int(5)

Fig. 9. Relation Database Workflow.

III. Result and Discussion

A. *Hospital*

The results of this hardware are the design that has been made according to the concept. The results can be seen in Fig. 10.

Fig. 10 shows the result of the hardware design, as seen on the image there is 3 hardware, which each hardware will be used in three cities and three different hospitals. The hardware also contains an LCD display that shows information about the hospital lab name, patient name, patient id, and queue number.

1) Doctor: On the Doctor’s main page will display the patient's ID to be filled by a doctor by scanning/inputting the QR Code that the patient has. The doctor's main page can be seen in Fig. 11.

On this action form page, there is information about the patient's personal data and there is an action form that will be filled out by the doctor after conducting an examination to the patient, also displaying the doctor's actions which can be seen in Fig. 12.

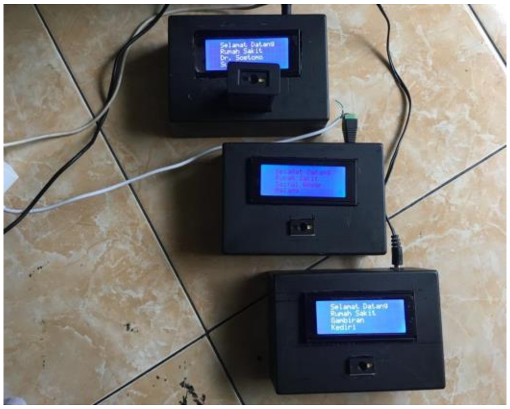
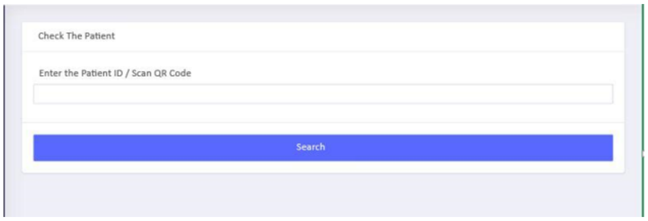
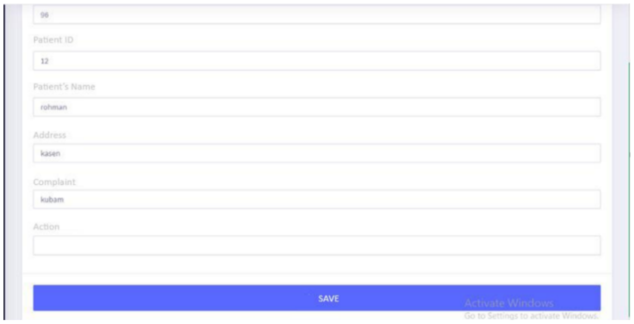


Fig. 10. Hardware Results.


Check The Patient

Enter the Patient ID / Scan QR Code

Fig. 11. Doctor Main Page.


96

Patient ID: 12

Patient’s Name: rohman

Address: kasen

Complaint: kubam

Action:

Fig. 12. Action Form Page/Patient Medical Records.

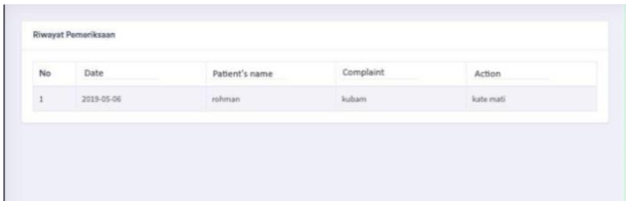
On the examination history page, there is an information display of patients who have been examined by a doctor. In the page, there is also information about the examination date, patient name, complaints, and actions given by the doctor. The check history page can be seen in Fig. 13.

Status menu page displays the names of patients who have registered for the doctor's examination. The status page also contains an information table consisting of the registration date column, order number, queue, patient name, complaints, and patient’s presence status. The status menu page can be seen in Fig. 14.

B. *User*

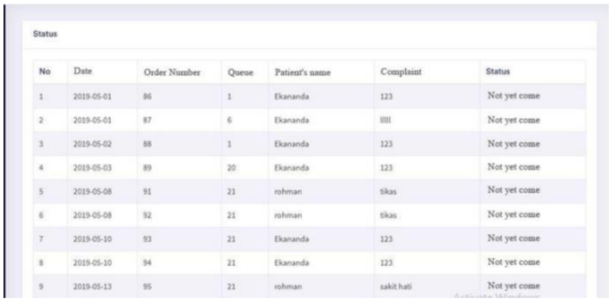
This login display is the initial display when the patient or doctor will access the website. In this login display, there are two inputs, namely, "Username" and "Password". To enter/log in the patient/doctor must enter both inputs. Before logging in, patients must first have an account by registering first. And the display of the login page can be seen in Fig. 15.

This registration page is a display when the patient will create a new account. In this registration view there are several inputs such as name, address, username, e-mail, and password where the patient will enter to enter the login page. The appearance of the registration page can be seen in Fig. 16.


Riwayat Pemeriksaan

No	Date	Patient's name	Complaint	Action
1	2019-05-06	rohman	kubam	kate mati

Fig. 13. Examination History Page.


Status

No	Date	Order Number	Queue	Patient's name	Complaint	Status
1	2019-05-01	86	1	Ekananda	123	Not yet come
2	2019-05-01	87	6	Ekananda	lllll	Not yet come
3	2019-05-02	88	1	Ekananda	123	Not yet come
4	2019-05-03	89	20	Ekananda	123	Not yet come
5	2019-05-08	91	21	rohman	tikas	Not yet come
6	2019-05-08	92	21	rohman	tikas	Not yet come
7	2019-05-10	93	21	Ekananda	123	Not yet come
8	2019-05-10	94	21	Ekananda	123	Not yet come
9	2019-05-13	95	21	rohman	sakit hati	Not yet come

Fig. 14. Status Menu Page.

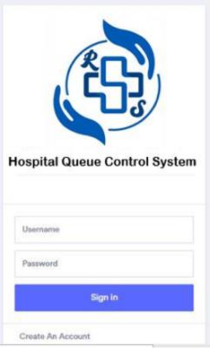
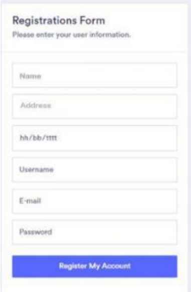


Fig. 15. Login Page.



Registrations Form
Please enter your user information.

Name
Address
hh/bb/tttt
Username
E-mail
Password

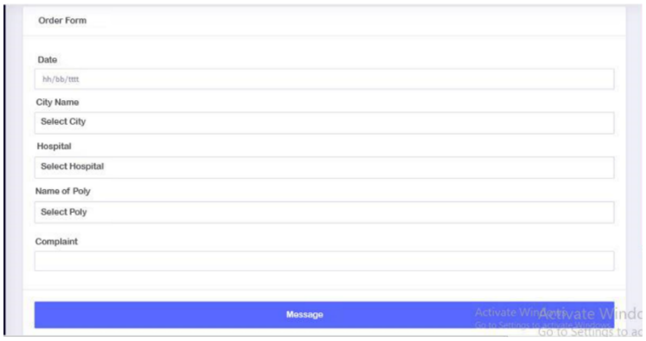
Fig. 16. Registration Page.

This order form page is used by patients to place an order by filling out a form consisting of Date, City, Hospital, hospital lab, and Complaints. On the first form the patient will choose the date first and then the patient chooses the city that the patient wants, where the system will display the hospital according to the city chosen and the hospital lab available at the hospital. After that the patient will enter a complaint that the patient will input according to the complaint of the disease, then the patient clicks on the message to go to the next process. The order form page can be seen in Fig. 17.

The display of the doctor's schedule page shows several names of doctors available at the hospital on that date. This page also provides information to patients to find out which doctor will serve the examination at the hospital. Then the patient clicks the message for the order process. The doctor's schedule page can be seen in Fig. 18.

This page, display the QR Code that will be obtained by the patient to verify arrival at the hospital. This QR Code can be stored by patients in the form of PDF. In this QR Code file, there is information such as the name of the hospital, the name of the hospital lab, order number, queue number and estimated time. This information is to make it easier for patients when doing treatment/examination at the hospital. The QR Code display can be seen in Fig. 19 and fill in the detailed information in the QR Code in PDF format can be seen in Fig. 20.

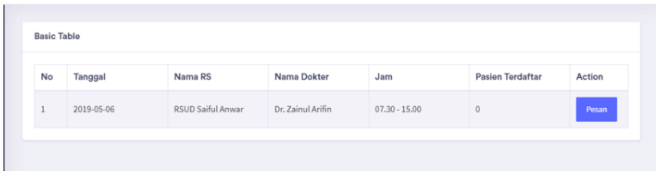
This map page shows the location of the hospital, which is useful to facilitate patients to go to the hospital, and the display of the maps page is shown in Fig. 21.



Order Form

Date
hh/bb/tttt
City Name
Select City
Hospital
Select Hospital
Name of Poly
Select Poly
Complaint

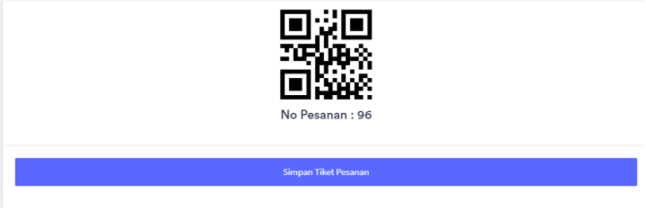
Fig. 17. Order Form.



Basic Table

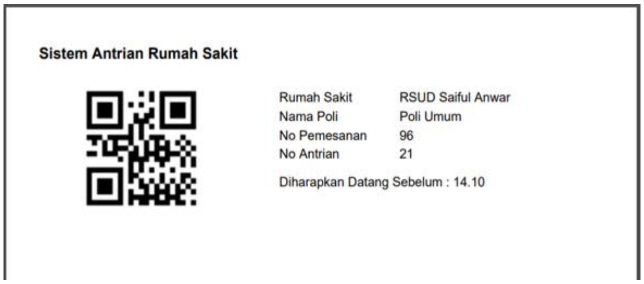
No	Tanggal	Nama RS	Nama Dokter	Jam	Pasien Terdaftar	Action
1	2019-05-06	RSUD Saiful Anwar	Dr. Zainul Arifin	07.30 - 15.00	0	Pesan

Fig. 18. Doctor Schedule Page.



No Pesanan : 96

Fig. 19. QR Code Page.

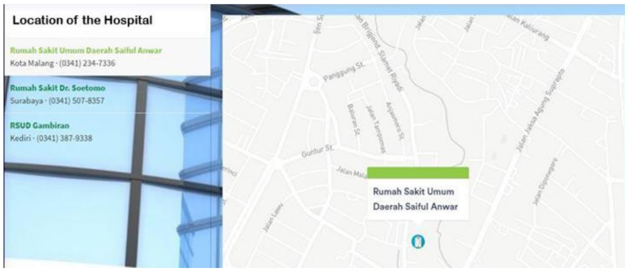


Sistem Antrian Rumah Sakit

Rumah Sakit	RSUD Saiful Anwar
Nama Poli	Poli Umum
No Pemesanan	96
No Antrian	21

Diharapkan Datang Sebelum : 14.10

Fig. 20. Maps in QR Code.



Location of the Hospital

Rumah Sakit Umum Daerah Saiful Anwar
Kota Malang · (0341) 234-7336

Rumah Sakit Dr. Soetomo
Surabaya · (0341) 507-8357

RSUD Gambiran
Kediri · (0341) 387-9338

Fig. 21. Maps.

IV. Conclusion

After measuring, testing and analyzing the data, it can be concluded that:

1) By using a website, we can make reservations in the hospital be done anywhere by patients which automatically produces a quick response code (QR Code).

2) For hospital bookings online, it will automatically be integrated so that patients can make hospital bookings listed on the website.

3) The use of integrated queue control systems in hospitals can help manage patients' personal data and medical records.

4) The quick response code (QR Code) can save the order ID that is used to verify the patient's arrival.

V. Future Work

This queue control system still has advantages and disadvantages, so the suggestions for this queue development system are as follows:

1) To develop the hospital registration or booking system, it can be developed on an Android application with a sound model and patient photos.

2) Need to feature hospital services and facilities when the patient's QR Code has been accumulated on Android.

References

[1] P.-F. Tsai and F.-M. Lin, “An Application of Multi-Attribute Value Theory to Patient-Bed Assignment in Hospital Admission Management: an Empirical Study,” J. Healthc. Eng., vol. 5, no. 4, pp. 439–456, 2014.

[2] S. M. Sohrevardi, A. Shafiei, and S. S. Mirzania, “Intravenous Immunoglobulin : A Drug Utilization Review at Shahid Sadoughi Hospital in Yazd,” no. 1, 2014.

[3] Ian Brown and Andrew Smale, “Management of Medical technology–Case study of a Major Acute Hospital” 29th Annual International Conference of the IEEE Engineeting in Medicine and Biology Society, 2007.

[4] “Intravenous Immunoglobulin : A Drug Utilization Review at Shahid Sadoughi Hospital in Yazd,” no. 1, 2014.

[5] S. M. Mousavi, F. Pourshariati, G. Rajabi, and M. Letafatnejad, “Waiting Time to Receive Healthcare Services and Factors Affecting It : Case Study in a University Hospital,” vol. 1, no. 2, 2017.

[6] Q. Su, X. Yao, P. Su, J. Shi, Y. Zhu, and L. Xue, “Hospital registration process reengineering using simulation method,” J. Healthc. Eng., vol. 1, no. 1, pp. 67–82, 2010.

[7] A. M. H. Pardede et al., “Framework For Patient Service Queue System For Decision Support System on Smart Health Care,” Int. J. Eng. Technol., vol. 7, no. 2.13, pp. 337–340, 2018H. N. Abed and C. Science, “Robust and Secured Image Steganography using LSB and Encryption with QR Code improvement of the amount and security of transmitting information , the secrecy of data method were proposed for information security . In cryptography, the encrypted message was,” vol. 9, no. 2, pp. 1–9, 2017.

[8] F. P. Mardiah and M. H. Basri, “The Analysis of Appointment System t o Reduce Outpatient Waiting Time at Indonesia’s Public Hospital,” Hum. Resour. Manag. Res., vol. 3, no. 1, pp. 27–33, 2013.

[9] Nutchanad Taveerad and Sartid Vongpradhip, “Development of Color QR Code for Increasing Capacity”, 11th International Conference on Signal-Image Technology & Internet-based, 2015, DOI :10.1109/SITIS.2015.42.

[10] P. Y. Lin and Y. H. Chen, “High payload secret hiding technology for QR codes,” Eurasip J. Image Video Process., vol. 2017, no. 1, 2017.

[11] Muming Li, Peng Cao, Liuping Feng, Lifang Yu, Jianbo Chen and Jing Wang, “The Research of QR code image correction based on image gray feature” First International Conference on Electronics Instrumentation & Information Systems (EIIS), 2017, DOI : 10.1109/EIIS.2017.8298583.

[12] Vassilya Uzun, “QR-Code Based Hospital Systems for Healtcare in Turkey”, IEEE 40th Annual Computer Software and Applications Conference (COMPSAC), volume 2 (2016), DOI: 10.1109/ COMPSAC.2016.173.

[13] D. A. Martillano, R. G. Reyes, I. R. B. Miranda, and K. L. C. Diaz, “Android-Based Smart Power Outlet Switching Device Using ESP8266 Enabled WiFi Module,” vol. 6, no. 1, pp. 61–65, 2018.

Artificial Immune-based Algorithm for Academic Leadership Assessment

Hamidah Jantan[1], Nur Hamizah Syafiqah Che Azemi[2]
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA (UiTM) Terengganu
Kuala Terengganu, Terengganu
Malaysia

Zulaiha Ali Othman[3]
Center of Artificial Intelligence Technology
Faculty of Information Science and Technology
Univsersiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

***Abstract*—Artificial Immune-based algorithm is inspired by the biological immune system as computational intelligence approach in data analysis. Negative selection algorithm is derived from immune-based algorithm's family that used to recognize the pattern's changes perform by the gene detectors in complementary state. Due to the self-recognition ability, this algorithm is widely used to recognize the abnormal data or non-self especially for fault diagnosis, pattern recognition, network security etc. In this study, the self-recognition performance proposed by the negative selection algorithm been considered as a potential technique in classifying employee's competency. Assessing the employee's performance in organization is an important task for human resource management people to identify the right candidate in job promotion assessment. Thus, this study attempts to propose an immune-based model in assessing academic leadership performance. There are three phases involved in experimental phase i.e. data acquisition and preparation; model development; and analysis and evaluation. The data consists of academic leadership proficiency was prepared as data-set for learning and detection processes. Several experiments were conducted using cross validation process on different model to identify the most accurate model. Therefore, the accuracy of NS classifier is considered acceptable enough for this academic leadership assessment case study. For enhancement, other immune-based algorithm or bio-inspired algorithms, such as genetic algorithm, particle swam optimization, ant colony optimization would also be considered as a potential algorithm for performance assessment.**

Keywords—Immune-based algorithm; negative selection algorithm; academic leadership; performance assessment

I. Introduction

In biological immune system, the learning process is through the evolution of distinguishing between our body's own cell and foreign antigen. This immune system behavior proposed a new paradigm in computational intelligence for data analysis [1]. Negative selection algorithm in Artificial Immune System (AIS) technique uses immunological principle by recognizing and categorizing the self on non-self in complementary basis for all body cells. The first negative selection algorithm was introduced by Forrest to identify information affected by the biological based infection and that was transformed to machine learning context [2]. This algorithm is mainly used to detect changes of pattern's behavior in complementary space performed by the gene detectors. The detectors are used to classify the new data whether it is self (normal) or non-self (ab-normal). This algorithm mostly used for anomaly detection by classifying the self and non-self's detectors such as in computer virus detection, network intrusion detection, fault diagnose, network security and many others area [3]. Besides that negative selection algorithm can also be used in pattern recognition area by matching the self or non-self's detectors in classifying the pattern.

The process of talent assessment in HR field involves highly on human decisions by managing employee's advancement through promotion exercises, those tasks are very subjective, uncertain and difficult [4]. In HR field, employee's assessment process is a way for an employee to enhance his/her career path development. This process can be implemented by using various approaches and it also depends on the employee's competency criteria based on their job specification. In higher learning institution, academic leadership competency refers to academic leadership ability such as in teaching and supervision, research and publication, expertise and contribution to university or society etc.

Due to the ability of Negative selection algorithm for classification in previous studies [1-3], this paper attempts to apply this algorithm for talent assessment as part of talent management task in Human Resource (HR) field by classifying the selected criteria for leadership assessment. In this study, negative selection algorithm is used as a potential method in assessing an employee for promotion based on biological immune system behavior. The paper is organized as follows. Section 2 describes immune-based algorithm, negative selection algorithm and academic leadership in higher learning institution. Section 3 explains the research method and Section 4 discusses on result of the experiments. Section 5 presents the conclusion and future directions for academic leadership assessment using immune-based system approach.

II. Related Work

A. Artificial Immune-Based Algorithm

Immune-based algorithm in computational intelligence using immunological principles inspired by biological immune system in the living organism for data analysis and known as Artificial Immune System (AIS) algorithms [1]. These algorithms are proposed from behavior of various types of natural immune systems. Clonal Se-lection Algorithm, Immune Networks Algorithm, Negative Selection Algorithm and

This study is funded by Fundamental Re-search Grant Scheme (FRGS) (600-RMI/FRGS 5/3 (3/2016).

Dendritic Cell Algorithm are among the algorithms that derived from AIS algorithm's family [5]. As an example, negative selection algorithm using immunological principle by recognizing cells within the body and categorize those cells as self on non-self in complementary basis which is this approach can be used to categorize or classify the important element in data analysis [3].

B. Negative Selection Algorithm

The Negative Selection (NS) algorithm is inspired by discrimination of the self and non-self through behavior observed in the mammalian learned in their immune system as shown in Fig. 1. Negative selection aims to provide tolerance for self-cells that deals with immune system's ability to detect unknown antigens while not re-acting to the self-cell. In biological immune system process for T-cell maturation, if a T-cell in thymus recognized any self-cell, it is eliminated, and then immune functionality is performed [2]. The NS behavior in natural immune system was gives a new direction for other Artificial Immune System (AIS) algorithms development.

This scenario has similarity with artificial immune system technique where the NS algorithm generates detector set by eliminating any detector candidate that matches with elements from a group of self-samples [5]. The job of T-cells will be performed by the population of detectors that is represented in the fixed length binary strings. A simple rule is used to compare bits in two such strings that has been selected from the population of detectors to check whether matching has occurred or not. This matching process is similar with natural mammals' immune system in matching between a lymphocyte and antigen. A sample of unlabelled data from certain area of problem domain will be trained to generate the set of detectors. Then, this detector will used to determine whether exist any new hidden data points belonging to the same sub-area or not.

NS algorithm uses the same principal where the detector has ability to recognise the pattern in small sub-area of the problem. This algorithm consists of two main steps as shown in Fig. 2. The first step is to generate a set of detectors to represent the unmatched string comparing with a predetermined substring as a memory cell. The string-matching rule usually applied on partial of the string, because it can be extremely rare that the generated random strings exactly match with the predetermined substring, even though the strings are small. There are several matching-rules that can be used in matching process as affinity measurement and it also depends on the data representation. For real data representation, R-Contiguous, R-chunk, Hamming Distance, Hamming-shape-space are commonly used matching-rule for matching process. The string data representation usually uses Euclidean Distance, Manhattan Distance and Makowski Distance. Next, the second step for NS algorithm is to perform comparing process between new data and the detector. If a detector recognized a change has occurred in string pattern, this will demonstrate the anomaly state exists in original data [6-7]. This approach might look simple and easy to execute, but it is an effective method especially when dealing with a small set of detectors strings as it will produce very high probability of recognizing the changes in the existing data.

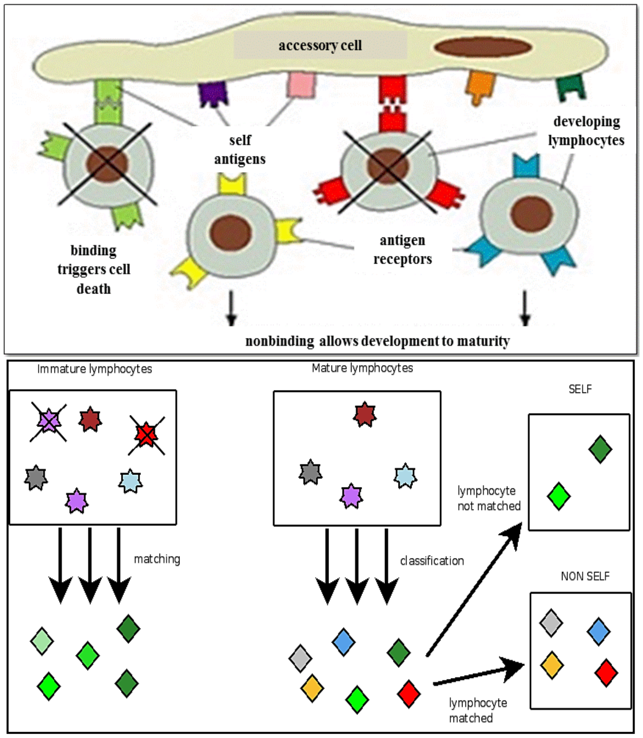


Fig. 1. Negative Selection Process in Vertebrate Immune System.

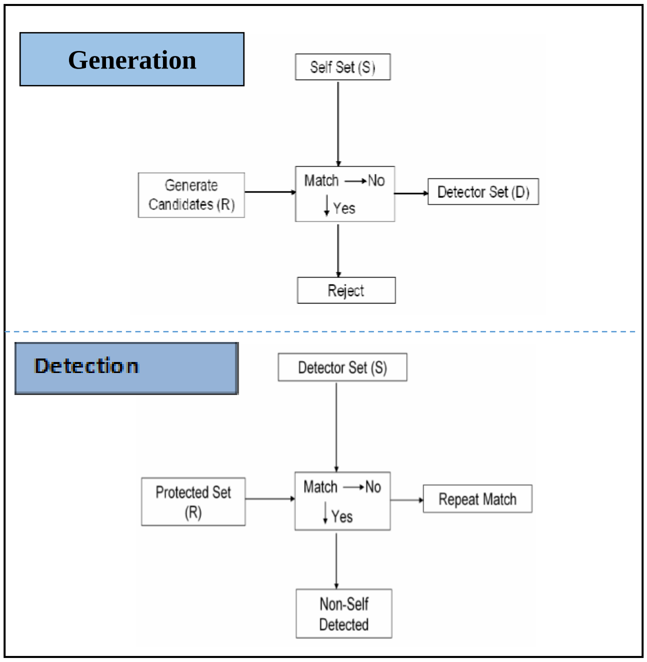


Fig. 2. Fundamental of Negative Selection Algorithm.

Recently, there are many areas applied NS algorithm to solve their problem and mainly it used anomaly detection by classifying self or non-self as detectors such as in computer virus detection, network security, network intrusion detection, fault diagnose, pattern recognition, business, etc. Table I shows several examples of application for specific purpose using negative selection algorithm technique for solving their problems [3].

TABLE. I. NEGATIVE SELECTION ALGORITHM APPLICATIONS

Application Area	**Issues/Problem domain**
Fault Diagnose	Aircraft Fault Detection [8] Disturbance Detection for Optimal Database Storage in Electrical Distribution Systems [9] Spam Detection [10] Fault Detection in Refrigeration Systems [11]
Network Security Detection	Network Intrusion Detection [12] Traffic Intrusion Detection [13]
Pattern Recognition	Image classification [14] Music Feature Recognition [15]
Classification / Selection	Merger and Acquisition Target Identification [16] Patient classification [17]

C. Academic Leadership Assessment in Higher Learning Institution

In higher learning institution, activities such as teaching, supervision, research, publication, service to university and service to the society are among the academic duties. These duties play an important role as competency measurement in managing academic leadership development [18]. Nevertheless, the emphasis and scope of these duties also depends on other factors such as university's direction, seniority and specialization, and academic appointment [19]. As an academia, teaching and supervising is about educating students in the specific field of study. In many cases, they also need to enhance their understanding especially when it involve with technology advancement. Besides that, they also must do academic research which is it can be extreme, inspiring, and worthwhile.

Academic research involves numerous exercises such as getting funding, knowledge sharing, knowledge transfer and many others besides exploring new discoveries in their field. They must spend their time in comprising for financing their research by preparing a good research proposal to attract interest from grant provider. In academic research, proposing experimental paper to reporting research discoveries and presenting this finding in any scientific conferences play an important role towards body of knowledge contribution. As researcher, there are enormous tasks that need to be carried out and experiment to conduct. Nevertheless, most of re-searchers are committed, patience and hard work in their duties since they love to do that [20]. In evaluating the competency of academic leadership in higher learning institution, all these aspects should be taken into consideration. Several individual factors that can be mapped with academic leadership context to produce academic leadership talent are shown in Fig. 3.

Nowadays, academic talent marketplace is highly competitive to determine the current and future direction for higher learning institution [21]. Academic talent is recognized by assessing academic leadership abilities that are needed to ensure the development of excellence in education-based institution. Academic leadership is an 'intellectual leadership' that focuses on the development of leading ideas and the establishment of new academic directions by giving the narrow and compartmentalized forms of knowledge as an academia. A higher learning institution needs to be able to identify the potential academic talent to ensure the competitive advantage for the institution will be achieved. The issues that associated with recruitment and retirement of academic talent is considered as the key of long-term success and competitiveness in higher learning institution. Hence, this study attempts to apply the Immune-based algorithm as an approach to classify the academic talent through academic leadership criterion. There are several researches had been conducted on this issue using soft computing and data mining techniques and they are proven successful [22, 23]. However, the evolutionary computation and bio-inspired algorithm such as Genetic Algorithm, Ant Colony Optimization and Artificial Immune System (AIS) has not attracted researchers in this area.

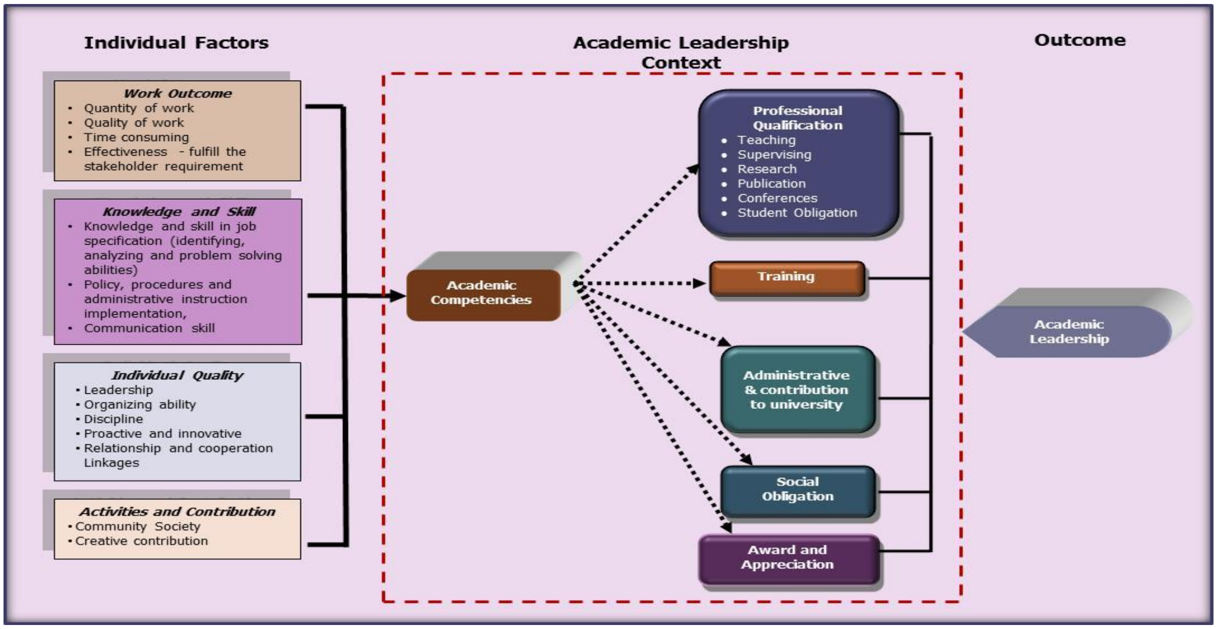

Individual Factors
Academic Leadership Context
Outcome
Work Outcome
• Quantity of work
• Quality of work
• Time consuming
• Effectiveness - fulfill the stakeholder requirement
Knowledge and Skill
• Knowledge and skill in job specification (identifying, analyzing and problem solving abilities)
• Policy, procedures and administrative instruction implementation,
• Communication skill
Individual Quality
• Leadership
• Organizing ability
• Discipline
• Proactive and innovative
• Relationship and cooperation Linkages
Activities and Contribution
• Community Society
• Creative contribution
Academic Competencies
Professional Qualification
• Teaching
• Supervising
• Research
• Publication
• Conferences
• Student Obligation
Training
Administrative & contribution to university
Social Obligation
Award and Appreciation
Academic Leadership

Fig. 3. Academic Leadership Mapping Criteria in Higher Learning.

III. Research Method

The experiment was conducted on NS algorithm as selected algorithm from immune-based algorithm as a case study in academic leadership assessment. The data was collected based on academic leadership criteria used in academic promotion evaluation. There are three phases involved in experimental phase i.e. data acquisition and preparation; model development; and analysis and evaluation as shown in Table II. The gathering information and data acquisition in the first phase focus on preparing the datasets for training process. The data was obtained from academic promotion evaluation results as sample of dataset. The selected attributes of this dataset are demographic and the evaluation assessment result from academic promotion criteria that represents the academic duties such as teaching, supervision, research, publication, consultation, service to university and leadership as shown in Table III.

The implementation of NS algorithm consists of two parts which is the learning process on the training dataset and the classification process to match the detector with the new candidates to determine the accuracy of NS classifier. The highest accuracy of classifier will be selected and applied to the new data for classification purposes. In the second phase, NS algorithm has been applied that involved three stages i.e. self-data definition, detector generation and detector matching is determined by affinity measurement. The self-data was produced randomly as the candidate that will be used in the training process as the first stage of implementation. The second stage is training phase or learning (generation) and testing phase or apply to new data (detection) that follow the basic process in negative selection algorithm, as shown in Fig. 4. Fig. 5 shows the model development framework for negative selection algorithm. Sample of training and testing dataset is shown in Fig. 6.

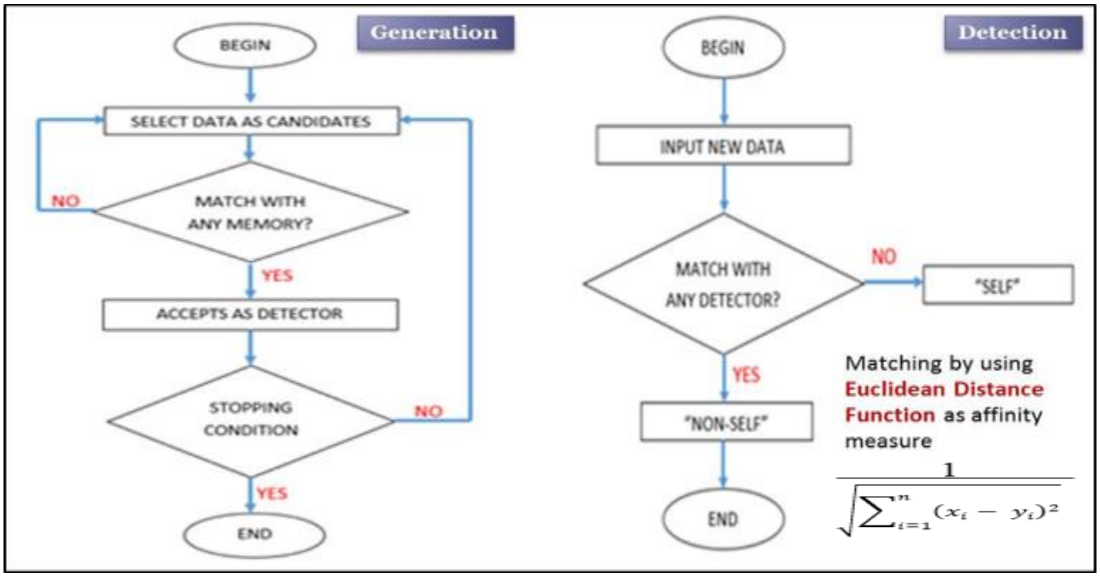
The matching process between candidate and initial memory cell that randomly generated will be executed. This process will recognize whether the candidate is self of non-self. Then, those candidates that match with self-data will be discard and would not be considered as detector. The detector for non-self will be used to detect and recognize anomaly or negative using matching process in detection process. The matching process is intent to know whether the new data match or not with the data in memory cell by calculate their affinity measure. The Euclidean distance function is used as affinity measurement for the matching process. This function will calculate the distance be-tween two focus points and compare with the threshold value to recognize the normal (positive) or abnormal (negative) data. This measurement method will help the process of filtering the non-self or negative cell.

TABLE. II. Experiment Phase Description

Phase	Description
Data-set	Academic leadership dataset collected from academic promotion evaluation data.
Model Development	**Phase 1: Generation** •Attributes – 7 attributes •Number of data – 80 •Method– Random generator **Phase 2: Detection /Matching rule** •Method– Euclidean distance based for real valued representation
Model Analysis	10-fold cross validation

TABLE. III. Data Description

Criteria	Attributes
Demographic	Year Gender Grade Promotion
Academic Leadership Criteria	Teaching and supervision Research and publication Consultation and/or Expertise Conference Participation Service University/Community Academic award Leadership and personal attitude



$$\frac{1}{\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}}$$

Fig. 4. Learning Process in Negative Selection Algorithm.

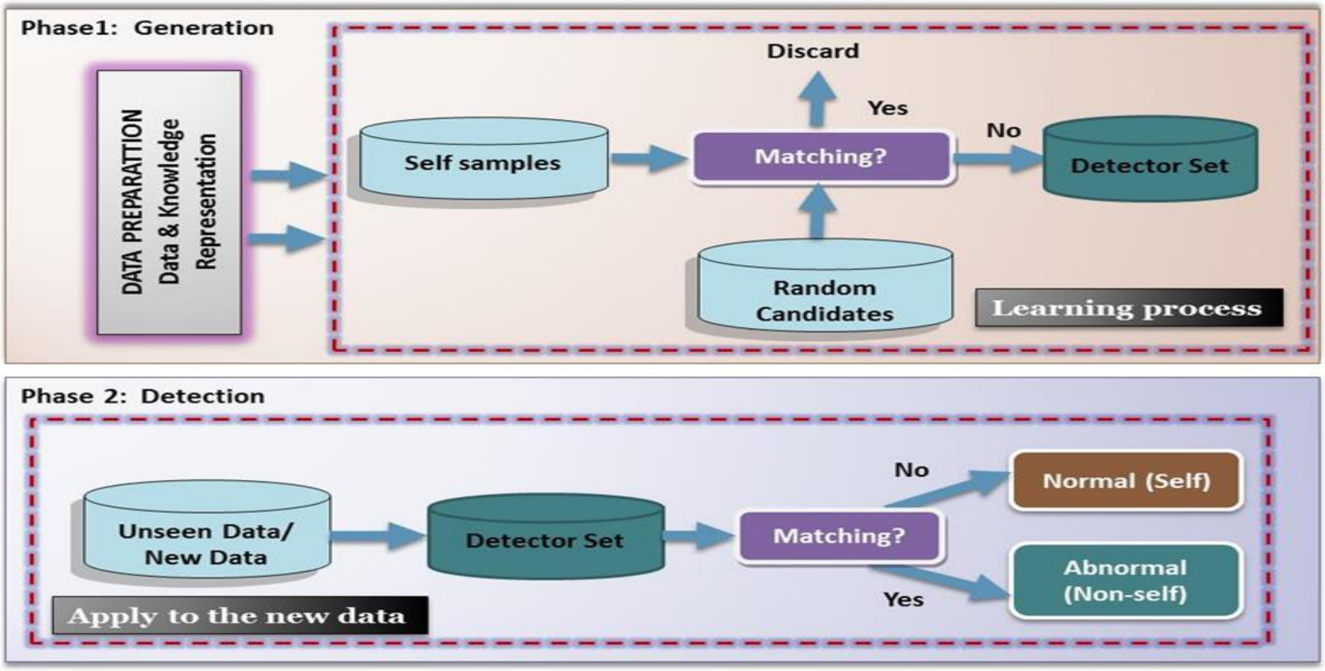
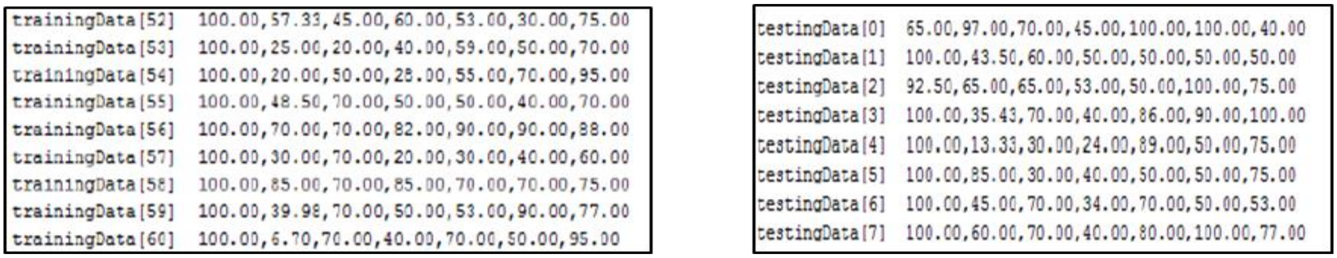


Fig. 5. Negative Selection Model Development Framework.



```
trainingData[52]   100.00,57.33,45.00,60.00,53.00,30.00,75.00
trainingData[53]   100.00,25.00,20.00,40.00,59.00,50.00,70.00
trainingData[54]   100.00,20.00,50.00,28.00,55.00,70.00,95.00
trainingData[55]   100.00,48.50,70.00,50.00,50.00,40.00,70.00
trainingData[56]   100.00,70.00,70.00,82.00,90.00,90.00,88.00
trainingData[57]   100.00,30.00,70.00,20.00,30.00,40.00,60.00
trainingData[58]   100.00,85.00,70.00,85.00,70.00,70.00,75.00
trainingData[59]   100.00,39.98,70.00,50.00,53.00,90.00,77.00
trainingData[60]   100.00,6.70,70.00,40.00,70.00,50.00,95.00
```

```
testingData[0]  65.00,97.00,70.00,45.00,100.00,100.00,40.00
testingData[1]  100.00,43.50,60.00,50.00,50.00,50.00,50.00
testingData[2]  92.50,65.00,65.00,53.00,50.00,100.00,75.00
testingData[3]  100.00,35.43,70.00,40.00,86.00,90.00,100.00
testingData[4]  100.00,13.33,30.00,24.00,89.00,50.00,75.00
testingData[5]  100.00,85.00,30.00,40.00,50.00,50.00,75.00
testingData[6]  100.00,45.00,70.00,34.00,70.00,50.00,53.00
testingData[7]  100.00,60.00,70.00,40.00,80.00,100.00,77.00
```

Fig. 6. Sample of Training and Testing Datasets.

IV. Result and Discussion

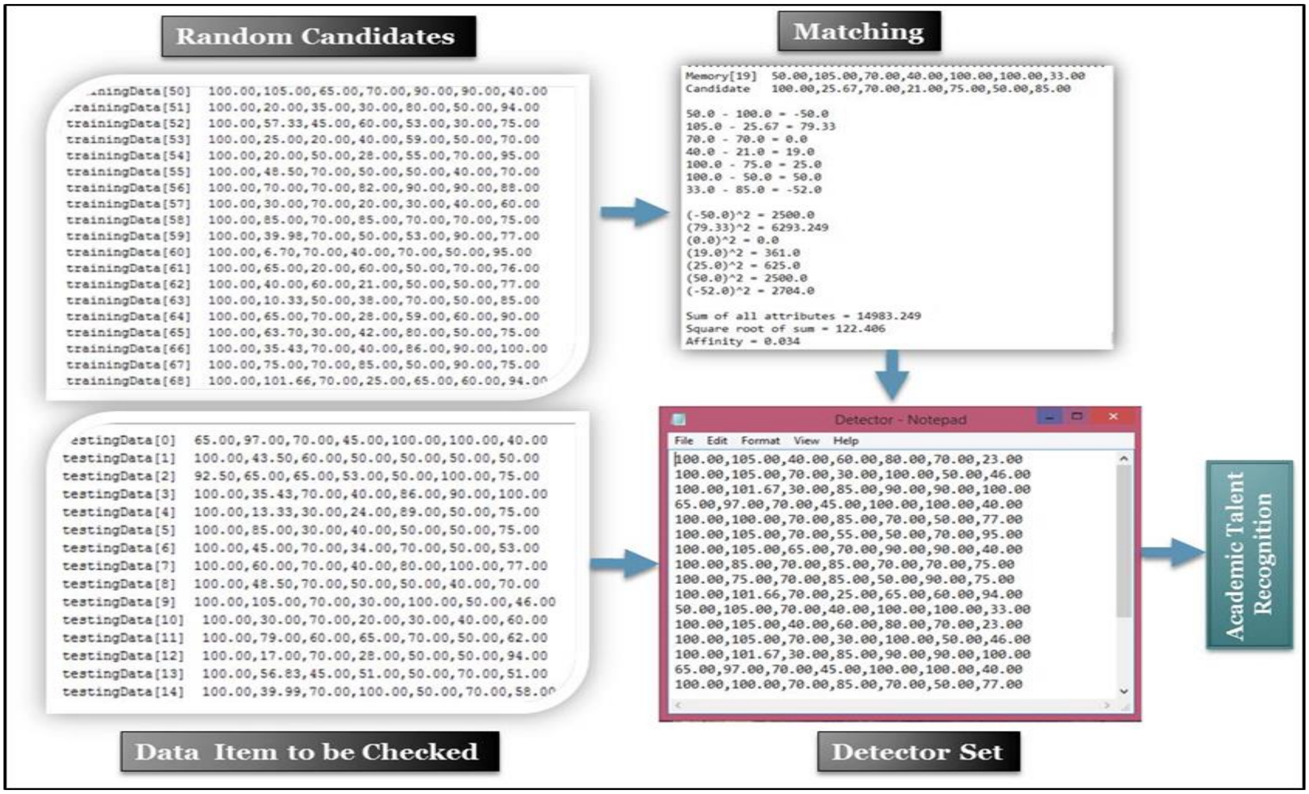
In this paper, the proposed negative selection classifier was developed through generation and detection processes that produce a set of detectors that are considered as non-self or negative selection classifier. Fig. 7 shows the process involved in this study that contains the generation and matching process to produce the set of detectors. The sample of negative selection detector as NS model produced in the learning proses is shown in Fig. 8. The accuracy of NS classifier was determined by 10-fold cross validation model in learning process using selected data set as mentioned in Table IV.

The accuracy of the model determined through matching between the candidates and set of identified detectors would be classified as negative, abnormal or non-self-cell. The result shows that 90:10 and 70:30 models produced highest accuracy. Besides that, the average of accuracy for NS algorithm is 85.86% and this would be considered as good and acceptable enough for classification. In model construction phase, the accuracy of the model is very important to produce an accurate result for classification and prediction.

As an example, result from evaluation phase for evaluating the proposed NS classifier to determine the right academia in promotion exercise is shown in Fig. 9. The result shows that NS algorithm can also be considered as potential approach for academic leadership assessment.

TABLE. IV. Negative Selection Model Analysis

10 Fold Cross Validation Model	**Accuracy**	**Description (Average)**
90:10	100.00	↑ (Potential Model)
80:20	66.67	↓
70:30	100.00	↑ (Potential Model)
60:40	80.00	↓
50:50	77.78	↓
40:60	90.00	↑ (Potential Model)
30:70	83.33	↓
20:80	91.67	↑ (Potential Model)
10: 90	83.33	↓



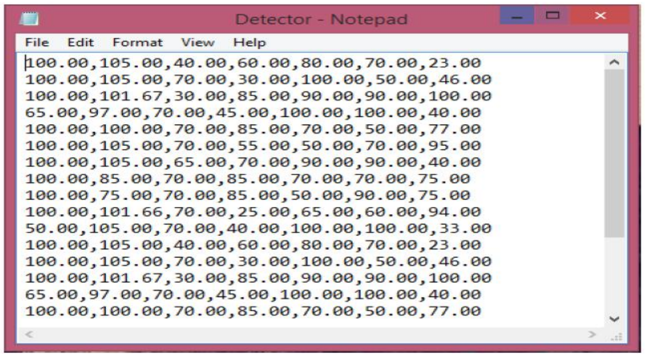
```
…ningData[50]   100.00,105.00,65.00,70.00,90.00,90.00,40.00
…rainingData[51] 100.00,20.00,35.00,30.00,80.00,50.00,94.00
trainingData[52] 100.00,57.33,45.00,60.00,53.00,30.00,75.00
trainingData[53] 100.00,25.00,20.00,40.00,59.00,50.00,70.00
trainingData[54] 100.00,20.00,50.00,28.00,55.00,70.00,95.00
trainingData[55] 100.00,48.50,70.00,50.00,50.00,40.00,70.00
trainingData[56] 100.00,70.00,70.00,82.00,90.00,90.00,88.00
trainingData[57] 100.00,30.00,70.00,20.00,30.00,40.00,60.00
trainingData[58] 100.00,85.00,70.00,85.00,70.00,70.00,75.00
trainingData[59] 100.00,39.98,70.00,50.00,53.00,90.00,77.00
trainingData[60] 100.00,6.70,70.00,40.00,70.00,50.00,95.00
trainingData[61] 100.00,65.00,20.00,60.00,50.00,70.00,76.00
trainingData[62] 100.00,40.00,60.00,21.00,50.00,50.00,77.00
trainingData[63] 100.00,10.33,50.00,38.00,70.00,50.00,85.00
trainingData[64] 100.00,65.00,70.00,28.00,59.00,60.00,90.00
trainingData[65] 100.00,63.70,30.00,42.00,80.00,50.00,75.00
trainingData[66] 100.00,35.43,70.00,40.00,86.00,90.00,100.00
trainingData[67] 100.00,75.00,70.00,85.00,50.00,90.00,75.00
trainingData[68] 100.00,101.66,70.00,25.00,65.00,60.00,94.0
```

```
Memory[19]   50.00,105.00,70.00,40.00,100.00,100.00,33.00
Candidate    100.00,25.67,70.00,21.00,75.00,50.00,85.00

50.0 - 100.0 = -50.0
105.0 - 25.67 = 79.33
70.0 - 70.0 = 0.0
40.0 - 21.0 = 19.0
100.0 - 75.0 = 25.0
100.0 - 50.0 = 50.0
33.0 - 85.0 = -52.0

(-50.0)^2 = 2500.0
(79.33)^2 = 6293.249
(0.0)^2 = 0.0
(19.0)^2 = 361.0
(25.0)^2 = 625.0
(50.0)^2 = 2500.0
(-52.0)^2 = 2704.0

Sum of all attributes = 14983.249
Square root of sum = 122.406
Affinity = 0.034
```

```
…estingData[0]   65.00,97.00,70.00,45.00,100.00,100.00,40.00
testingData[1]   100.00,43.50,60.00,50.00,50.00,50.00,50.00
testingData[2]   92.50,65.00,65.00,53.00,50.00,100.00,75.00
testingData[3]   100.00,35.43,70.00,40.00,86.00,90.00,100.00
testingData[4]   100.00,13.33,30.00,24.00,89.00,50.00,75.00
testingData[5]   100.00,85.00,30.00,40.00,50.00,50.00,75.00
testingData[6]   100.00,45.00,70.00,34.00,70.00,50.00,53.00
testingData[7]   100.00,60.00,70.00,40.00,80.00,100.00,77.00
testingData[8]   100.00,48.50,70.00,50.00,50.00,40.00,70.00
testingData[9]   100.00,105.00,70.00,30.00,100.00,50.00,46.00
testingData[10]   100.00,30.00,70.00,20.00,30.00,40.00,60.00
testingData[11]   100.00,79.00,60.00,65.00,70.00,50.00,62.00
testingData[12]   100.00,17.00,70.00,28.00,50.00,50.00,94.00
testingData[13]   100.00,56.83,45.00,51.00,50.00,70.00,51.00
testingData[14]   100.00,39.99,70.00,100.00,50.00,70.00,58.0
```

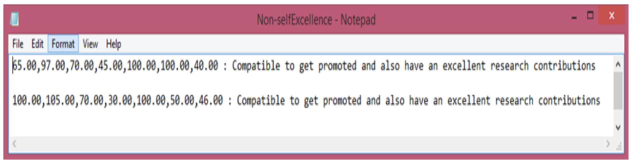
```
100.00,105.00,40.00,60.00,80.00,70.00,23.00
100.00,105.00,70.00,30.00,100.00,50.00,46.00
100.00,101.67,30.00,85.00,90.00,90.00,100.00
65.00,97.00,70.00,45.00,100.00,100.00,40.00
100.00,100.00,70.00,85.00,70.00,50.00,77.00
100.00,105.00,70.00,55.00,50.00,70.00,95.00
100.00,105.00,65.00,70.00,90.00,90.00,40.00
100.00,85.00,70.00,85.00,70.00,70.00,75.00
100.00,75.00,70.00,85.00,50.00,90.00,75.00
100.00,101.66,70.00,25.00,65.00,60.00,94.00
50.00,105.00,70.00,40.00,100.00,100.00,33.00
100.00,105.00,40.00,60.00,80.00,70.00,23.00
100.00,105.00,70.00,30.00,100.00,50.00,46.00
100.00,101.67,30.00,85.00,90.00,90.00,100.00
65.00,97.00,70.00,45.00,100.00,100.00,40.00
100.00,100.00,70.00,85.00,70.00,50.00,77.00
```

Fig. 7. Negative Selection Analysis Process.



```
100.00,105.00,40.00,60.00,80.00,70.00,23.00
100.00,105.00,70.00,30.00,100.00,50.00,46.00
100.00,101.67,30.00,85.00,90.00,90.00,100.00
65.00,97.00,70.00,45.00,100.00,100.00,40.00
100.00,100.00,70.00,85.00,70.00,50.00,77.00
100.00,105.00,70.00,55.00,50.00,70.00,95.00
100.00,105.00,65.00,70.00,90.00,90.00,40.00
100.00,85.00,70.00,85.00,70.00,70.00,75.00
100.00,75.00,70.00,85.00,50.00,90.00,75.00
100.00,101.66,70.00,25.00,65.00,60.00,94.00
50.00,105.00,70.00,40.00,100.00,100.00,33.00
100.00,105.00,40.00,60.00,80.00,70.00,23.00
100.00,105.00,70.00,30.00,100.00,50.00,46.00
100.00,101.67,30.00,85.00,90.00,90.00,100.00
65.00,97.00,70.00,45.00,100.00,100.00,40.00
100.00,100.00,70.00,85.00,70.00,50.00,77.00
```

Fig. 8. Negative Selection Classifier /Detector.



```
65.00,97.00,70.00,45.00,100.00,100.00,40.00 : Compatible to get promoted and also have an excellent research contributions

100.00,105.00,70.00,30.00,100.00,50.00,46.00 : Compatible to get promoted and also have an excellent research contributions
```

Fig. 9. Model Evaluation for Academic Leadership Assessment.

V. Conclusion

This paper proposed NS algorithm from immune-based system approach as classifier for data analysis in academic leadership assessment as a case study. However, the proposed method only focuses on research and publication datasets for assessment. In future work, the proposed method could be compared to other similar method especially for common comparison framework on other academic leadership criteria. Besides that, there are other immune-based algorithm such as Artificial Immune Network, Clonal Selection, and Dendritic Cell algorithms need to be explored as a comparative study. It would give a new understanding in this field specially to select the most accurate result as produced by the algorithm in academic leadership assessment. Finally, this result will give a new direction of applying immune-based algorithm in human re-source activities for data analysis.

Acknowledgment

This research has been supported by Ministry of Education (MOE) to Universiti Tekonologi MARA (UiTM) for Fundamental Re-search Grant Scheme (FRGS) (600-RMI/FRGS 5/3 (3/2016).

References

[1] Timmis, J, Neal, M, & Hunt, J (2000), An Artificial Immune System for Data Analysis. Biosystems, Vol. 55, No. (1-3), pp.143-150. doi: 10.1016/S0303-2647(99)00092-1.

[2] Forrest S, Perelson A S, & Allen L. Self-nonself Discrimination in a Computer (1994), IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, pp.202-212, doi: 10.1109/RISP.1994.296580.

[3] Wei Y G, Zheng D L., & Wang Y (2004), Negative Selection Algorithm and Its Application in Anomaly Detection. Third International Conference on Machine Learning and Cybemetics, Shanghai, China, pp. 2910-2913, doi: 10.1109/ICMLC.2004.1378529.

[4] Jantan H, Hamdan A R, & Othman Z A (2011), Data Mining Classification Techniques for Human Talent Forecasting: InTech, pp.1173-1178.

[5] Elsayed S A M, Ammar R A, & Rajasekaran S (2012), Artificial Immune Systems: Models, Applications, and Challenges. 27th Annual ACM Symposium on Applied Computing Trento, Italy, pp.256-258.

[6] Ji Z, & Dasgupta D (2007), Revisiting Negative Selection Algorithms. Evolutionary Computation, Vol. 15, No. 2, pp.223-251. doi: 10.1162/evco.2007.15.2.223.

[7] Soam S S, Khan F, Bhasker B, & Mishra B N (2011), Identification of MHC Class II binders/ non-binders using Negative Selection Algorithm. Journal of Bioinformatics and Sequence Analysis, Vol. 5, No. 2, pp.16-24. doi: 10.5897/JBSA11.023.

[8] Dasgupta D, KrishnaKumar K, Wong D, & Berry M (2004), Negative Selection Algorithm for Aircraft Fault Detection. International Conference on Artificial Immune Systems, Vol. 3239, pp.1-13, doi: 10.1007/978-3-540-30220-9_1.

[9] Lima F P A, Lotufo A D P, & Minussi C R (2014), Disturbance Detection for Optimal Database Storage in Electrical Distribution Systems using Artificial Immune Systems with Negative Selection. Electric Power Systems Research, Vol. 109, pp.54-62. doi: 10.1016/j.epsr.2013.12.010.

[10] Idris I, & Selamat A (2011), Negative selection algorithm in artificial immune system for spam detection. 5th Malaysian Conference in Software Engineering (MySEC), Johor Bahru, Malaysia, pp.379-382, doi: 10.1109/MySEC.2011.6140701.

[11] Taylor D W & Corne D W (2003), An Investigation of the Negative Selection Algorithm for Fault Detection in Refrigeration Systems. International Conference on Artificial Immune Systems, Vol. 2787, pp.34-45, doi: 10.1007/978-3-540-45192-1_4.

[12] Kim J & Bentley P J (2001), An Evaluation of Negative Selection in An Artificial Immune System for Network Intrusion Detection. Genetic and Evolutionary Computation Conference (GECCO).

[13] Hooks D, Yuan X, Roy K, Esterline & Joaquin Hernandez J (2018), Applying Artificial Immune System for Intrusion Detection. 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), Bamberg, Germany, doi: 10.1109/BigDataService.2018.00051.

[14] Bendiab E & Kholladi M K (2012), Insupervised Classification Based Negative Selection Algorithm. Courrier du Savoir, Vol. 14, pp.31-35

[15] Muda A K, Muda N A & Choo Y H (2017), Recognizing Music Features Pattern Using Modified Negative Selection Algorithm for Songs Genre Classification. Advances in Intelligent Systems and Computing, Vol. 734, pp.242-251.

[16] Paul S, Janecek A, Neto F B D L & Marwala T (2013), Applying the Negative Selection Algorithm for Merger and Acquisition Target Identification. BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence (BRICS-CCI & CBIC), Ipojuca, Brazil, doi: 10.1109/BRICS-CCI-CBIC.2013.107.

[17] Bölükbaş O & Harun Uğuz H (2018), Performance of Negative Selection Algorithms in Patient Detection and Classification. Nature-Inspired Intelligent Techniques for Solving Biomedical Engineering Problems, Vol. 14, pp.25-31.

[18] Tariquea I & S Schuler R (2010), Global talent management: Literature review, integrative framework, and suggestions for further research. Journal of World Business, Vol. 45, No. 2, pp.122-133. doi: 10.1016/j.jwb.2009.09.019.

[19] Ismail M & Rasdi R M (2008), Leadership in an Academic Career: Uncovering the Experience of Women Professors. International Studies in Educational Administration (Commonwealth Council for Educational Administration & Management (CCEAM)), Vol. 3, No. 36, pp.87-103.

[20] Vincent-Lancrin S (2006), What is Changing in Academic Research? Trends and Futures Scenarios. European Journal of Education, Vol. 41, No. 2. doi: 10.1111/j.1465-3435.2006.00255.x.

[21] Verhaegen P (2006), Academic talent: Quo vadis? Recruitment and retention of faculty in European business schools. Journal of Management Development, Vol. 24, No. 9.

[22] Chang J R, Cheng C H & Chen L S (2007), A Fuzzy-based Military Officer Performance Appraisal System. Applied Soft Computing, Vol. 7, No. 3, pp.936-945. doi: 10.1016/j.asoc.2006.03.003.

[23] Jantan H, Hamdan A R & Othman Z A(2009), Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application. World Academy of Science, Engineering and Technology, Vol. 3, No.2, pp.775-78.

Antennas of Circular Waveguides

Cusacani Guerrero[1], Julio Agapito[2]
Electronics and Telecommunication Department
Universidad Nacional Tecnológica de Lima Sur
Lima, Peru

Roman-Gonzalez Avid[3]
Aerospace Sciences and Health Research Laboratory
(INCAS-Lab), Universidad Nacional Tecnológica de Lima
Sur, Lima, Perú

***Abstract*—The design of the circular waveguide antenna is proposed for displacement reflector antennas. For them, we use the frequencies of operation so that our waveguide generates the mode, (Transversal Electric), resulting in a high impedance bandwidth. The results obtained from the radiation pattern of the fabricated antenna give excellent results according to the numerical data. Used as a primary feed-in compensation, reflector decreases cross-polarization.**

***Keywords*—Circular waveguide antenna; mode; microwave oven**

I. Introduction

The circular waveguide antenna is a conductive cylinder, through which the electromagnetic waves are transferred, radiating within it. The waveguides operate in the frequency range of (300 MHz and 30 GHz), being its manufacture mostly of metallic components. The transmission of signals in waveguides reduces the dissipation of energy. The signal goes along the guide, limiting its borders. Also, to no loss in the dielectric, because this is air. They are most often used in equipment such as radars, which need a rotating antenna and microwave.

The article shows a circular waveguide mode TE_{11}, as seen in Fig. 1. The antenna designed in the mode TE_{11}, that produces a cutoff frequency f_c, on which the antenna will operate. The first and second index of the modes refers to the azimuthal and radial variations, respectively. The simulated and measured results are finalized by comparing, demonstrating an excellent agreement. The numerical results are carried through the Ansoft HFSS program. The presented antenna is very compact, and the mode TE_{11} it reduces cross-polarization, being able to be used in reflecting antennas. This situation is so because the mode cancels or minimizes cross-polarization [1].

II. Methodology

The geometry of the circular waveguide antenna is shown in Fig. 1. The desired operating mode is the TE_{11}. The desired radius is r = 14mm, the distance between the dielectric and the metal is th = 2mm, the length of the waveguide is L = 70mm to meet the criteria of the cutoff frequency f_c 6.3 GHz. To simplify the understanding of each mode in the far field region, formulas are used, which will be mentioned below.

A. Equations

The circular waveguides are used in waveguides, which are a hollow conductor, which is filled with a dielectric material. The solution of the wave equation in cylindrical coordinates, for electric and magnetic fields, is done with maxwell equations for variable fields in time [2]:

$$\vec{\nabla}x\vec{H} = j\omega e\vec{E} \tag{1}$$

$$\vec{\nabla}x\vec{E} = -j\omega\mu\vec{H} \tag{2}$$

$$\nabla^2\vec{H} = \gamma^2\vec{H} \tag{3}$$

$$\nabla^2\vec{E} = \gamma^2\vec{E} \tag{4}$$

One uses equations 3 and 4, and one gets the Laplacian wave function [3]:

$$\nabla^2\psi = \gamma^2\psi \tag{5}$$

From where one obtains the Helmholtz scalar equation in cylindrical coordinates to solve it by the method of separation of variables [4]:

$$\psi = \psi(r,\phi,z) \tag{6}$$

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\psi}{\partial r}\right)+\frac{1}{r^2}\frac{\partial^2\psi}{\partial\phi^2}+\frac{\partial^2\psi}{\partial z^2}=\gamma^2\psi \tag{7}$$

One simplifies by separating variables:

$$\Psi = R(r)\Phi(\phi)Z(z) \tag{8}$$

Substitute in z and divide in equation (7) for the functionψ, resulting:

$$\frac{1}{rR}\frac{d}{dr}\left(r\frac{dR}{dr}\right)+\frac{1}{r^2\Phi}\frac{d^2\Phi}{d\phi^2}+\frac{1}{Z}\frac{d^2Z}{dz^2}=\gamma^2 \tag{9}$$

Being the sum of the terms in terms of z:

$$\frac{d^2Z}{dz^2}=\gamma^2 z \tag{10}$$

Being γ^2, the propagation constant in air, being the solution of the homogeneous linear partial differential equation:

$$Z(z) = Ae^{-\gamma_g z}+Be^{\gamma_g z} \tag{11}$$

One replaces equation (8) in (9):

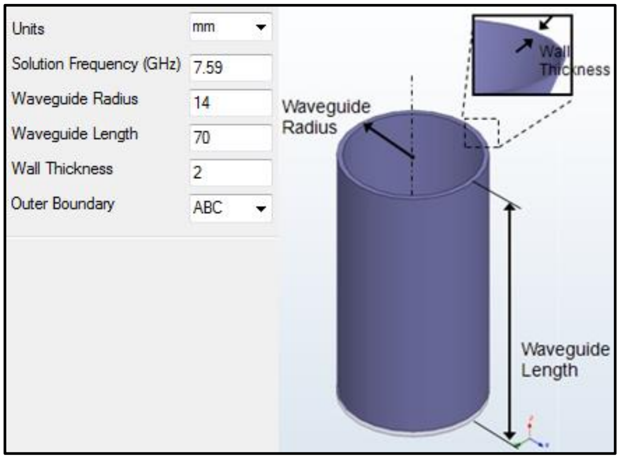
$$\frac{r}{R}\frac{d}{dr}\left(r\frac{dR}{dr}\right)+\frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2}-\left(\gamma^2-\gamma_g^2\right)r^2=0 \tag{12}$$

Now one operates according to z:

$$\frac{\partial^2\Phi}{\partial\phi^2}=-k_\phi^2\Phi \tag{13}$$

Being your solution:

$$\Phi(\phi) = A\sin(k_\phi\phi)+B\cos(k_\phi\phi) \tag{14}$$



Units	mm
Solution Frequency (GHz)	7.59
Waveguide Radius	14
Waveguide Length	70
Wall Thickness	2
Outer Boundary	ABC

Fig. 1. Geometry of the Antenna Operating in Mode TE_{11}.

One replaces in the equation (8):

$$r\frac{d}{dr}\left(r\frac{dR}{dr}\right)+\left[(k_c r)^2-k_\phi^2\right]R=0 \tag{15}$$

Equation (13) is a Bessel equation, where one has the cutoff wave number:

$$k_c^2=k^2-\beta^2 \tag{16}$$

Being the constant of propagation in the air:

$$\beta^2=\pm\sqrt{\omega^2\mu\varepsilon-k_c^2} \tag{17}$$

$$\gamma_g=\alpha_g+j\beta_g \tag{18}$$

Being the function of Bessel in function of R (r):

$$R(r)=C.J_n(k_c r)+D.Y_n(k_c r) \tag{19}$$

So $J_n(x)$ y $Y_n(x)$, are functions of bessel of first and second class of order n, being of the longitudinal magnetic field formed by the multiplication of R (r) and $\Phi(\phi)$, however, we do not use the second-class Bessel function, because when evaluated at the origin it takes an infinite value, because it does not reflect the electromagnetic field at the origin of the waveguide [5]:

$$\psi(r,\phi)=\left(A\sin\left(k_\phi\phi\right)+B\cos(k_\phi\phi)\right)J_n(k_c r) \tag{20}$$

Evaluating the electric field direction of ϕ when r = a:

$$\left(E_\phi(r,\phi)|_{r=a}\right)=0 \tag{21}$$

In order for this condition to be fulfilled, a cut wave number will be chosen that causes the derivative of the Bessel function to be canceled, so that there is a zero of the function when r = a. We will find this cut wave number, from the radius of the circular waveguide and also the cutoff frequency [5]:

$$k_c=\frac{\Phi'_{nm}}{a} \tag{22}$$

n: It's the order of Bessel's function

m: Ordinal number of zeros of the Bessel function

Being the cutting frequency of the circular waveguide:

$$f_{c_{nm}}=\frac{k_c}{2\pi a\sqrt{\mu\varepsilon}} \tag{23}$$

$$\beta_{nm}=\sqrt{k^2-k_c^2} \tag{24}$$

The impedance of the wave is [5]:

$$Z_{TE}=\frac{E_x}{H_y}=\frac{k\eta}{\beta} \tag{25}$$

B. Operation Tables

Below are the tables for the operating modes of our circular waveguide.

The first four first-class Bessel Functions are shown in Fig. 2, as mentioned in the second-class or Newman functions because it tends to infinity, and one will only be left with the first-class Bessel function J_n, of the graph [6].

The oscillation of Bessel's functions allows tabulating the arguments for which they are worth zero, that is, when the axis of the abscissa is crossed. These roots give rise to the frame for the Transversal Electric TE and Transversal Magnetic TM [7], modes shown in Table I and Table II, respectively.

For the handling of circular waveguides, standards such as the IEC system (Electronic Industry Association, United States) have been established, classifying the guides with letter C followed by a number, Table III provides data on these standards, such as Cutoff frequency of operating modes, such as hypothetical levels of dominant mode attenuation, for a reference frequency.

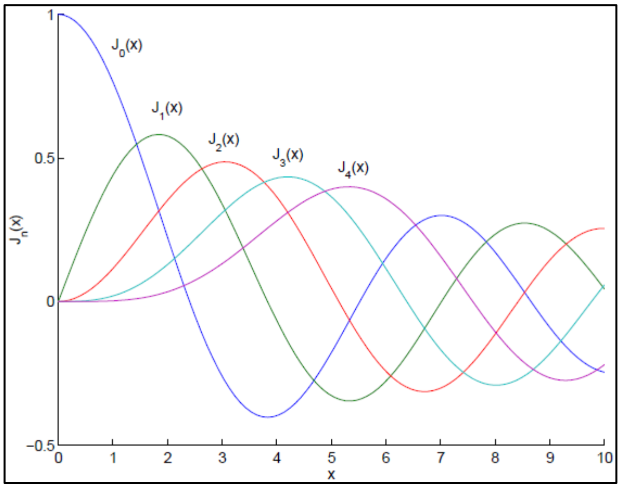


Fig. 2. Bessel Function Chart of First Class.

TABLE. I. ROOTS FOR TE MODES

n	n=1	n=2	n=3
m=0	3.832	7.016	10.174
m=1	1.841	5.331	8.536
m=2	3.054	6.706	9.970

TABLE. II. ROOTS FOR TM MODES

n	n=1	n=2	n=3
m=0	2.405	5.520	8.654
m=1	3.832	7.016	10.174
m=2	5.135	8.417	11.620

TABLE. III. LIST OF STANDARDS FOR CIRCULAR GUIDES IEC

Designación	Radio (mm)	Frecuencia de corte (GHz)		f (GHz)	Atenuación (dB/m)
		TE11	TM01		
C30	35.7	2.46	3.21	2.95	0.0184
C35	30.5	2.88	3.76	3.45	0.0233
C40	26.0	3.38	4.41	4.06	0.0297
C48	22.2	3.95	5.16	4.74	0.0375
C56	19.0	4.61	6.02	5.53	0.0473
C65	16.3	5.40	7.05	6.48	0.0599
C76	13.9	6.32	8.26	7.59	0.0759
C89	11.9	7.37	9.63	8.85	0.0956
C140	7.54	11.6	15.2	13.98	0.1893
C290	3.56	24.6	32.2	29.54	0.5834

III. RESULTS

Studies were conducted to determine the parameters of the waveguide antenna. Determining based on the measurements of the antenna and the equations.

A. *Theoretical Results*

To find the theoretical results we use the formulas mentioned above.

1) Based on the radius that is equal to r = 14mm, we can determine the cutting frequency f_c for the mode TE_{11},, being this f_c=6.3 GHz, as well as the f_c for the mode TM_{01}, it would f_c =8.2GHz. Obtaining the theoretical bandwidth of the difference of these, would be BW=1.9GHz. Finding these frequencies of the modes to work in the optimal frequency of 7.59 GHz, according to the IEC system, which provides these standards in Table I.

2) We find the wavelength of the dielectricλ = 0.0395, the wavelength of the group is $\lambda_g = 0.0708m$, with which we define the size of L. Obtaining also the phase of the mode $TE_{11}, \beta = 88.72\ rad/m.$

B. *Simulation Results*

The results of the simulation are obtained from the circular waveguide antenna obtained with the Ansoft HFSS program. Within the parameters that we will calculate first with the dimensions of our antenna will be the cutoff frequency f_c, that is obtained in Fig. 3. As it is observed there are two frequencies of cut for the two modes, in which the frequency 6.3 GHz for being the smaller one will be our frequency of cut below which the ones of more frequencies in our antenna.

Radiation diagrams are a graphic representation of the radiation properties of the antenna as a function of the different directions of space at a fixed distance. Express the electric field, in the far field area where the shape of the diagram is invariant as a function of distance. The following diagram shows the Cartesian and 3D polar radiation diagram in Fig. 4 and 5. Obtaining as a result both a directional radiation diagram, which means that we can cover more distant distances, but the effective beam amplitude decreases, what can not cover large areas.

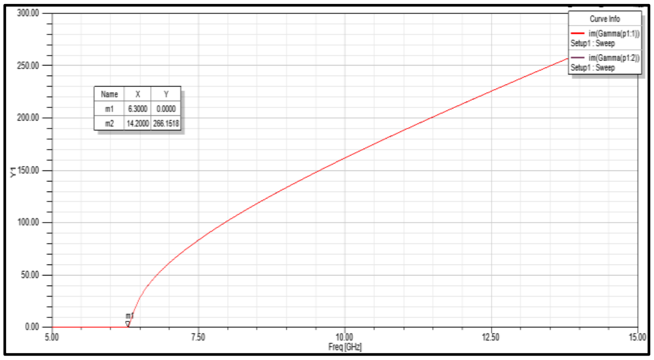


Fig. 3. Graph of the Diagram f_c in Ansoft HFSS.

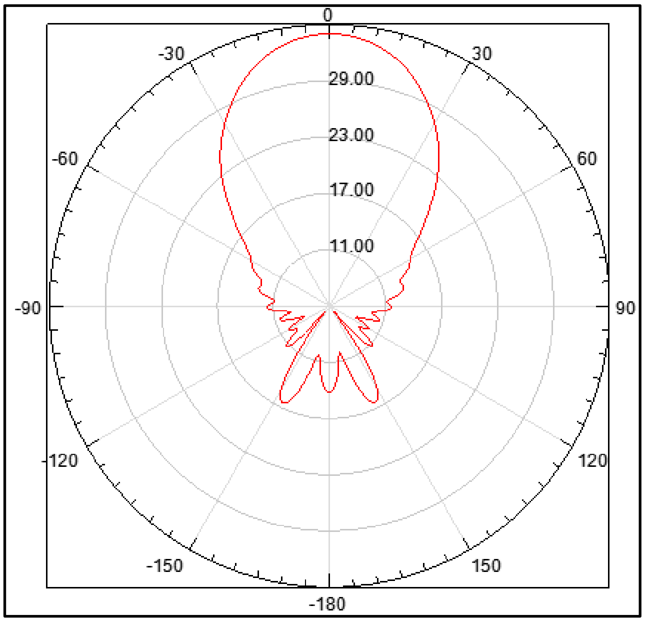


Fig. 4. Radiation Diagram in Cartesian Coordinates.

As it is observed the main lobe has an angle of 0º, while the rear lobe has an angle of 180º, finding the rear lobes between these two.

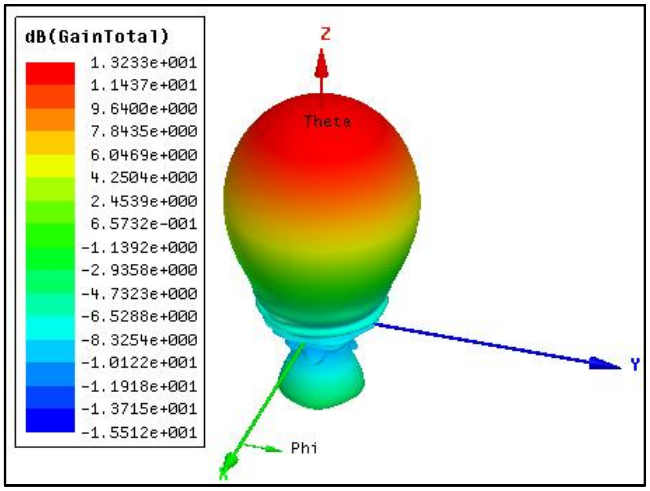


Fig. 5. 3D Radiation Diagram.

IV. DISCUSSION

It is concluded by the theoretical and simulated results obtained that the waveguide antenna is good for operating at microwave frequencies (300 MHz - 30GHz), cutting frequency f_c =6.3 GHz. Showing in the graph of Fig. 5, its radiation pattern is more directive, which reaches a greater distance, ideal for radio link communication. This polarization is due to the fact that the dominant pattern can be oriented in any direction of the z-axis of the guide, so that the two dominant modes can be transmitted at the same time, oriented their electric fields with a phase difference of 90 °. Compared to other antennas, the satellite dish has a high gain due to the shape of its radiation pattern that is more direct, unlike the waveguide antenna, which has a high average gain. Being among the advantages of waveguide antennas their circular or cross polarization, so that the reception of both vertical and horizontal polarization, is very convenient for satellite communication that requires using this type of polarization, in addition to being optimal for working in the microwave bands of C (4 to 8 GHz) and X (8 to 12 GHz).

REFERENCES

[1] Z.Allahgholi Pour and L. Shafai “A Ring Choke Excited Compact Dual-Mode Circular Waveguide Feed for Offset Reflector Antennas” IEEE Trans. Antennas Propag., VOL. 60, NO. 6,JUNE 2012, p. 1-2.

[2] Fawwaz T. Ulaby, Fundamentos de aplicaciones en Electromagnetismo, Quinta Edición., The University of michigan, 2007, pp.286–319.

[3] Erwin Kreyszig, Matemáticas Avanzadas para Ingeniería, Volumen 2, Cuarta edición, Ohio State University, Editorial Limusa, S.A. y Jhon Wiley & Sons (HK), Ltd., 2013, pp.59–123.

[4] Miguel Angel Fuentes Pascual, Representación en Matlab de modos resonantes en cavidades esféricas, Trabajo Fin de Grado presentado en la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universitat Politécnica de Valéncia para la obtención del Título de Graduado en Ingeniería de Tecnologías y Servicios de Telecomunicación, 2017, pp.10–31.

[5] Ebert Gabriel San Román Castillo,Patricia Raquel Castillo Araníbar, Manuel Gustavo Sotomayor Polar, Lee Victoria Gonzales Fuentes, Efraín Zenteno Bolaños, Aplicaciones y Teoría de Ingeniería de Microondas, 1 era ed., Iniciativa Latinoamericana de Libros de Texto Abiertos (LATin), 2014, pp. 24-39.

[6] Alfonso Zozaya, Ondas Guíadas, Curso Académico Sistemas Avanzados de Transmisión I, Unidad I, Aspectos Generales sobre Guías de Onda, Universidad de Antioquia, 2009, pp 26-35.

[7] Rodolfo Neri Vela, Líneas de Transmisión, Nueva edición complementada con presentaciones electrónicas y laboratorios virtuales interactivos realizados por Luis H. Porragas Beltrán, Primera Edición., Universidad Veracruzana Dirección General Editorial, 2013, pp.321–350.

Pedestrian Crossing Safety System at Traffic Lights based on Decision Tree Algorithm

Denny Hardiyanto[1], Muamar Rojali[4]
Department of Electrical Engineering
Institut Sains and Teknologi
AKPRIND Yogyakarta
Yogyakarta, Indonesia[1, 4]

Iswanto[2], Dyah Anggun Sartika[3]
Department of Electrical Engineering, Universitas[2]
Muhammadiyah Yogyakarta, Yogyakarta, Indonesia
Department of Computer Control Engineering, Politeknik[3]
Negeri Madiun, Madiun, Indonesia

***Abstract*—Pedestrians are one of the street users who have the right to get priority on security. Highway users such as vehicle drivers sometimes violate the traffic lights that is endanger pedestrians and make pedestrians feel insecure when crossing the street. Based on this problem, a tool is designed to provide a warning for the drivers or riders violating the traffic lights and prevent traffic accident by spraying water. The system is able to detect traffic violation based on changes in the value of the vehicle position on the stop line obtained from the Ultrasonic HC-SR04 sensor. When a violation is detected, a decision tree algorithm turns on the pump to spray water to the traffic violators as a deterrent effect. The results show that the vehicle located closest to the sensor has 94% precision, 88% recall and 85% accuracy, the vehicle located in the middle has 73% precision, 100% recall, and 75% accuracy, and the vehicle located furthest to the sensor has 75% precision, 100% recall and 80% accuracy.**

Keywords—Pedestrian safety; decision tree algorithm: traffic light; spraying water

I. Introduction

The pedestrians have the right to get the most priority security and comfort on the street. Motorcycle and car users sometimes endanger pedestrians by violating the traffic lights. A safety system to protect the pedestrians from those who violate the traffic lights is needed. Based on this problem, this paper presents a pedestrian safety system at the cross roads.

Some previous researchers have researched pedestrian safety. Jia examined pedestrian identification of planar objects using imaging of light field. Recognizing fake pedestrians on planar surfaces (2-D fake pedestrians) is an important and challenging task in the field of vision of the machine because of its various applications. The 2-D pedestrian recognition method was proposed based on light field imaging and support vector machines. This method could recognize 2-D fake pedestrians with only one sensor in one exposure [1]. Wang examined detection of pedestrians through the body parts of semantic and contextual information with DNN. The Estonian Pedestrian has increased in recent years, while the handling of complex occlusion and accurate localization is still the most important problem. The section networks and contexts consisted of three branches, namely, basic branch, section branch, and the context branch. This specifically uses two branches to detect pedestrians through semantic information on body parts and contextual information, respectively. By combining output from all branches, a strong complementary pedestrian detectors with lower error rates and higher localization accuracy was developed, especially for pedestrian occlusion [2].

Xu examined signalization of multi-level pedestrians at the large four-foot roundabout. Multi-level pedestrian signalization (MPS) was developed for roundabouts to balance pedestrian accessibility and vehicle mobility. With the right application of traffic control devices and traffic detection systems, the correct road was assigned to pedestrians and vehicles in three modes that were driven. Signals at adjacent crossings could operate independently or as groups in special mode of road rights to prioritize pedestrians or vehicles [3]. Zang investigated pedestrian detection with a scale-conscious localization policy. A major obstacle to pedestrian detection lied in the sharp decline in performance in the presence of small size pedestrians relatively far from the camera. An active pedestrian detector that explicitly operates in multilayer neuronal representations of input still images was developed. More specifically, convolutional neural nets, such as ResNet and R-CNN were exploited to provide a rich and discriminatory feature hierarchy of representations, as well as early pedestrian proposals [4].

An investigation on car brakes for the safety of pedestrians was examined by Avinash. The system used a proximity sensor to detect pedestrians. The sensor was processed by a microcontroller. The data processed by the microcontroller were then released to activate the Selenoid relay moving the brake to stop the car or motorbike [5]. A pedestrian flow calculated by using image processing was examined by Madhira. The system used a camera to capture the pedestrian images which were then processed by using minicomputers. A sensor camera was installed with several tilt angle positions to detect the pedestrians [6]. A FPGA-based traffic lights controller was examined by Kishore. The system consisted of an FPGA, a seven-segment display, a traffic lights control, a red LED, and a green LED. The FPGA was used to control the traffic lights timing [7]. A low-light pedestrian detection of RGB images using multi-modal knowledge distillation was examined by Kruthiventi. The system used an RGB camera processed by using a computer. Conventional and neural network algorithms were used to process the image data [8].

An effective detecting object from traffic camera videos was examined by Shi. The system consisted of cameras installed in the traffic lights. The fast RCNN algorithm and R-FCN method were used for object detection. The cameras installed in the traffic lights were processed by using computers with both algorithms [9]. Pedestrian Re-Identification Based on Image Enhancement Strategies and Over-fitting Solutions was investigated by Ding. The camera data were inserted into the computer to be processed by using several algorithms. The LHEMR algorithm was used to process unclear image data which were processed by using Fuzzy algorithms [10]. A Test for Smart Traffic lights Control on Pedestrian Cross on the street was examined by Tian. The system consisted of a light traffic control and a context-aware and intelligent system. The system was tested in real-like conditions on the street [11]. The Method of Pedestrian Detection Based on the YOLOv3 Model and Image Enhancement by Retinex was examined by Qu. The sensor used a camera whose data were processed by using a computer. The OpenCV python algorithm was used to process image data using the YOLOv3 Model [12].

Real-time Pedestrian Traffic Light Detection was examined by Ash. A camera was used in the sensor for object detection. The camera data were processed by computer using the R-CNN algorithm, KCF Tracker, and Tiny YOLO2 [13]. The Motion Detection Sensor Application for Monitoring Traffic Direction in Smart Street Lighting Systems was investigated by Tetervenoks. The system consisted of an infrared sensor used to detect moving objects. The data from the sensor were processed with an amplifier before inserted into the microcontroller [14]. Traffic lights scheduling for pedestrians and vehicles was examined by Zhang. The system consisted of traffic lights for pedestrians and vehicles. The traffic lights scheduling was modeled with a vehicle and pedestrian flow model. The Linear integer programming algorithm was used for scheduling [15]. A classification and introduction objects from a moving laser scanning point in the cloud of the road environment was examined by Lehtomaki. The system consisted of a laser radar that functions for input sensor from the system. The method was carried out by using pre-processing and removing ground and facades. Segmentation and segmentation classification were needed to obtain location points [16].

Referring to the previous researches, a pedestrian safety system has been examined. The previous system used camera sensor and laser sensor to detect the pedestrians. Microcomputers such as Raspberry and FPGA were used to process the data from the sensors. The purpose of this paper is to present a pedestrian security system when crossing the street. The system presented in this paper is different from the previous papers. The system uses Arduino microcontroller to process sensor data and activate spray when a traffic lights violation occurs.

II. Research Method

Pedestrian crossing safety system at traffic lights intersection is divided into two, namely, system design and pedestrian crossing safety system prototype.

1) System planning: Electronic and mechanical devices used in pedestrian crossing safety system are illustrated in Fig. 1. It can be seen that the system includes microcontroller devices using Arduino Uno [17]–[20], Relay, Ultrasonic Sensor, Power Supply, and DC pump. The control system used is a programmable Arduino Uno. An ultrasonic sensor is used to detect objects using ultrasonic waves. A relay is used as a switch to turn on actuators based on the signals obtained from the sensors, while a DC pumps is the actuator for spraying water.

The design of the program decision tree algorithm [21] for pedestrian crossing safety systems is shown by the flowchart in Fig. 2 [22]–[25]. The figure describes that the power supply turns on the device to activate the traffic light system. The ultrasonic sensor will be active when the traffic light turns red and gives an input signal when there is movement of an object in front of it. The Arduino processes the input signal and sends the signal to the relay. The relay serves as an on/off switch to run the DC pump to spray water as a deterrent effect on traffic violators.

2) System prototype: The design of the prototype using Sketch-up application can be seen in Fig. 3. The figure shows that the prototype size is 80 cm long, 40 cm wide, and 10 cm high. The prototype was made as similar as possible to the condition at the traffic light crossing. The study was conducted using toy car prototypes the detection objects of the ultrasonic sensor.

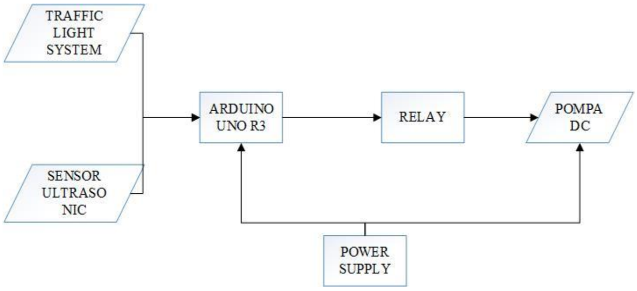


Fig. 1. Block Diagram of Pedestrian Security System.

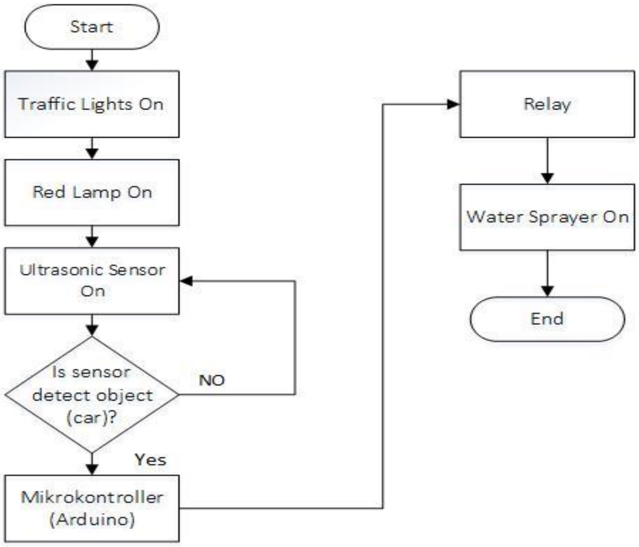


Fig. 2. The Working Principle of the Pedestrian Security.

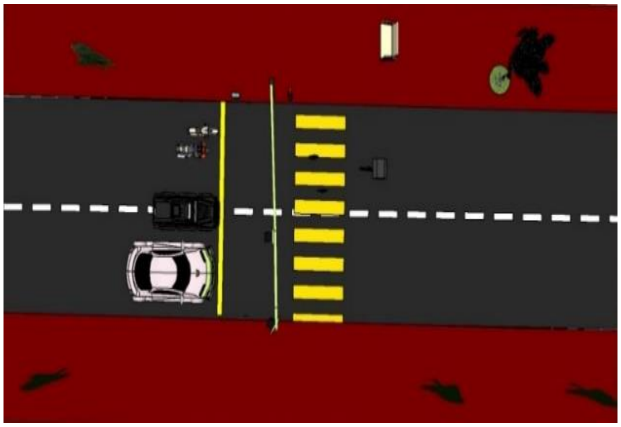


Fig. 3. Pedestrian Security System Prototype.

III. Discussion

Fig. 4 shows the test of pedestrian crossing safety system and the overall system. It can be seen that the pedestrian crossing safety system consists of three systems namely input system, control system and output system. The input system consists of a sonar sensor which is a proximity sensor. The control system is the system that processes the input, and then the microcontroller proceeds and releases it to turn on/off the relay. The output system is the system that activates the water pump to spray vehicle drivers or riders.

1) Change in sensor distance value: The test of the overall system was carried out to determine the safe distance set for the pedestrian safety system as shown in Table I. The table shows that that there are columns for the distance values set, columns for measurement values, and relay conditions. The distance values specified in this paper are 17 cm and 18 cm. The measurement test starts from the closest distance to the farthest. This paper presents the closest distance of 17 cm to the farthest distance of 22 cm. It can be seen that when the measurement distance is further than 4 cm, the system cannot detect make the relay OFF.

2) Changes in the position of the vehicle against the stop line: There were three conditions carried out in this test in which the vehicles was in position A, B, and C as shown in Fig. 5.

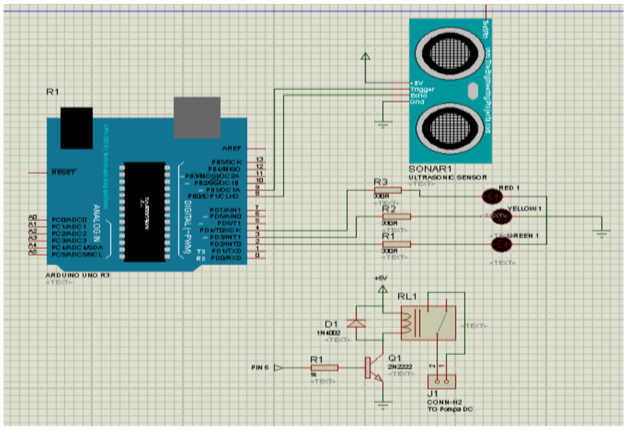


Fig. 4. Electronic Circuit of Pedestrian Crossing Safety Systems.

TABLE I. Values of Sensor Distance and Measurement Results

Data	Values set	Measurement values	Relay states
1	17 cm	17 cm	*ON*
2	17 cm	18 cm	*ON*
3	17 cm	19 cm	*ON*
4	17 cm	20 cm	*ON*
5	17 cm	21 cm	*OFF*
6	18 cm	18 cm	*ON*
7	18 cm	19 cm	*ON*
8	18 cm	20 cm	*ON*
9	18 cm	21 cm	*ON*
10	18 cm	22 cm	*OFF*

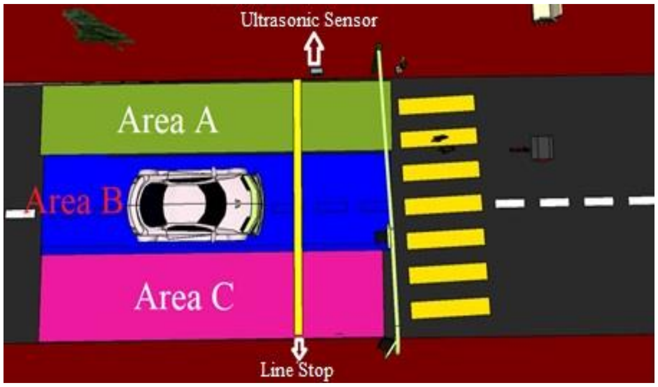


Fig. 5. Distribution of 3 Vehicle Positions (Area).

Fig. 5 shows the division of the three vehicle positions: Area A in green is closest to the sensor, Area B in blue is the middle position, and area C in pink is the farthest position to the sensor. In this experiment, the data were collected 20 times in area A, which is shown in Table II. The table is divided into columns for vehicle position, for the system and for the remark. The remark column presents TP for True Positive, TN for True Negative, and FN for False Negative, and FP for False Positive. It is seen in Table II that the negative value (-) indicates that the vehicle has not crossed yet the stop line, while the positive value (+) indicates that the vehicle has passed the stop line.

In this experiment, the data were collected 20 times in area B as shown in Table III. It can be seen that there are columns for the value of vehicle position, for the system and for the remark. The remark column shows TP for True Positive, TN for True Negative, FN for False Negative, and FP for False Positive. It can be seen in Table III that the negative value (-) indicates that the vehicle has not crossed yet the stop line, while the positive value (+) indicates the vehicle has passed the stop line. The correct value is ranging from +0.1 cm to +5 cm when the vehicle passes the stop line on which the system is ON (spraying water), and when the vehicle has not crossed the stop line or stops just on the stop line with the value less than 0 cm on which the system is OFF or water is not spraying.

In this experiment, data was collected 20 times in area B which is shown in Table IV. From the table, it can be seen that there are columns for the value of vehicle position, system column and description. In the description column, load TP which is True Positive, TN is True Negative, FN is False Negative, and FP is Positive False. In Table IV, you can see the negative value (-) in the numbers indicating that the vehicle has not crossed the stop line while the positive (+) value indicates the vehicle has passed the stop line. Correct value is when the vehicle passes the stop line which is worth +0.1 cm to +5 cm the system is ON (spraying water) and when the vehicle has not crossed the stop line or stops just above the stop line with a value less than 0 cm the system is OFF or water is not spraying. In Table IV, there are 5 errors that the system does not work correctly. This is influenced by the sensitivity of ultrasonic sensors when detecting vehicles.

TABLE II. SYSTEM STATES WHEN THE VEHICLE IS IN AREA A (NEAR THE SENSOR)

Data	Vehicle position values	System state	Remark
1	- 1 cm	*OFF*	*TP*
2	+ 1 cm	*ON*	*TP*
3	+ 2 cm	*ON*	*TP*
4	+0,4 cm	*OFF*	*FN*
5	+2,5 cm	*ON*	*TP*
6	+3 cm	*ON*	*TP*
7	+3,5 cm	*ON*	*TP*
8	+4 cm	*ON*	*TP*
9	+4,3 cm	*ON*	*TP*
10	+5 cm	*ON*	*TP*
11	+ 5,3 cm	*ON*	*FP*
12	+ 6 cm	*OFF*	*TP*
13	-1,5 cm	*OFF*	*TP*
14	-2 cm	*OFF*	*TP*
15	-2,5 cm	*OFF*	*TP*
16	+0,5 cm	*ON*	*TP*
17	-0,5 cm	*OFF*	*TP*
18	No vehicles	*OFF*	*TN*
19	+0,1 cm	*OFF*	*FN*
20	-0,1 cm	*OFF*	*TP*

TABLE III. SYSTEM STATES WHEN THE VEHICLE IS IN AREA B

Data	Vehicle position values	System state	Remark
1	- 1 cm	*OFF*	*TP*
2	+ 1 cm	*ON*	*TP*
3	+ 5 cm	*ON*	*TP*
4	-0,1 cm	*ON*	*FP*
5	+0,5 cm	*ON*	*TP*
6	+0,1 cm	*ON*	*TP*
7	+3,5 cm	*ON*	*TP*
8	+4 cm	*ON*	*TP*
9	+4,5 cm	*ON*	*TP*
10	+5 cm	*ON*	*TP*
11	+ 6 cm	*ON*	*FP*
12	+ 6,5 cm	*ON*	*FP*
13	-0,5 cm	*OFF*	*TP*
14	-2 cm	*OFF*	*TP*
15	-2,5 cm	*OFF*	*TP*
16	-0,3 cm	*ON*	*FP*
17	+3 cm	*ON*	*TP*
18	0 cm	*ON*	*FP*
19	+7 cm	*OFF*	*TP*
20	No vehicles	*OFF*	*TN*

TABLE IV. SYSTEM STATES WHEN THE VEHICLE IS IN AREA C

Data	Vehicle position values	System state	Remark
1	- 1 cm	*OFF*	*TP*
2	- 0,4 cm	*ON*	*FP*
3	+ 1 cm	*ON*	*TP*
4	0 cm	*ON*	*FP*
5	+5 cm	*ON*	*TP*
6	+3 cm	*ON*	*TP*
7	+3,5 cm	*ON*	*TP*
8	+4 cm	*ON*	*TP*
9	+4,5 cm	*ON*	*TP*
10	+5 cm	*ON*	*TP*
11	+ 6 cm	*ON*	*TP*
12	+ 6,5 cm	*ON*	*TP*
13	-1,5 cm	*OFF*	*TP*
14	-2 cm	*OFF*	*TP*
15	-2,5 cm	*OFF*	*TP*
16	+0,5 cm	*ON*	*TP*
17	No vehicles	*OFF*	*TN*
18	+7,3 cm	*ON*	*FP*
19	-0,1 cm	*ON*	*FP*
20	+0,1 cm	*ON*	*TP*

The values of precision, recall, and accuracy for each area are summarized in Table V. The table shows that area A has the highest accuracy and precision. This is because area A is the area closest to the sensor, while the lowest accuracy value is in area B due to the sensor sensitivity in detecting the presence of the vehicles.

TABLE V. VALUES OF PRECISION, RECALL AND ACCURACY OF SYSTEM

Area	Precision	Recall	Accuracy
A	94 %	*88 %*	*85 %*
B	73 %	*100 %*	*75 %*
C	75 %	*100 %*	*80 %*

IV. CONCLUSION

Based on the results of testing and discussion, it can be concluded that a pedestrian crossing safety system has been realized at the traffic lights intersection. The placement of the sensor is very influential on reading values and systems. This system is able to detect a vehicle closest to the sensor or in Area A with 85% accuracy, 88% recall and 94% precision. The system is able to detect a vehicle in area B with 75% accuracy, 100% recall and 73% precision. The system is able to detect a vehicle in area C with 80% accuracy, 100% recall and 75% precision. After the traffic violator has been detected by the system, the tool will activate the water pump. The time for water spray is ± 5 seconds.

REFERENCES

[1] C. Jia, F. Shi, Y. Zhao, M. Zhao, Z. Wang, and S. Chen, “Identification of Pedestrians From Confused Planar Objects Using Light Field Imaging,” IEEE Access, vol. 6, pp. 39375–39384, 2018.

[2] S. Wang, J. Cheng, H. Liu, F. Wang, and H. Zhou, “Pedestrian Detection via Body Part Semantic and Contextual Information With DNN,” IEEE Trans. Multimed., vol. 20, no. 11, pp. 3148–3159, Nov. 2018.

[3] H. Xu, K. Zhang, Q. Zheng, and R. Yao, “Multi-level pedestrian signalisation at large four-leg roundabouts,” IET Intell. Transp. Syst., vol. 12, no. 8, pp. 838–850, 2018.

[4] X. Zhang, L. Cheng, B. Li, and H.-M. Hu, “Too Far to See? Not Really!—Pedestrian Detection With Scale-Aware Localization Policy,” IEEE Trans. Image Process., vol. 27, no. 8, pp. 3703–3715, Aug. 2018.

[5] R. Avinash, J. Niresh, V. H. Kumar, and S. Neelakrishnan, “Investigation of pedestrian collision avoidance with auto brake,” in 2017 Recent Developments in Control, Automation & Power Engineering (RDCAPE), 2017, vol. 3, pp. 477–481.

[6] K. Madhira and A. Shukla, “Pedestrian flow counter using image processing,” in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 1911–1915.

[7] S. V. Kishore, V. Sreeja, V. Gupta, V. Videesha, I. B. K. Raju, and K. M. Rao, “FPGA based traffic light controller,” in 2017 International Conference on Trends in Electronics and Informatics (ICEI), 2017, pp. 469–475.

[8] S. S. S. Kruthiventi, P. Sahay, and R. Biswal, “Low-light pedestrian detection from RGB images using multi-modal knowledge distillation,” in 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 4207–4211.

[9] H. Shi, Z. Liu, Y. Fan, X. Wang, and T. Huang, “Effective object detection from traffic camera videos,” in 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2017, pp. 1–5.

[10] Y. ding, “Pedestrian Re-identification Based on Image Enhancement and Over-fitting Solution Strategies,” in 2018 5th International Conference on Systems and Informatics (ICSAI), 2018, no. Icsai, pp. 745–750.

[11] Y. Tian, X. Ma, C. Luo, and Y. Zhang, “A Testbed for Intelligent Control of Traffic Lights at Pedestrian Crossings on a Road,” in 2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP), 2018, pp. 1–6.

[12] H. Qu, T. Yuan, Z. Sheng, and Y. Zhang, “A Pedestrian Detection Method Based on YOLOv3 Model and Image Enhanced by Retinex,” in 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2018, no. 2016, pp. 1–5.

[13] R. Ash, D. Ofri, J. Brokman, I. Friedman, and Y. Moshe, “Real-time Pedestrian Traffic Light Detection,” in 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE), 2018, pp. 1–5.

[14] O. Tetervenoks, A. Avotins, P. Apse-Apsitis, L. R. Adrian, and R. Vilums, “Movement Detection Sensor Application for Traffic Direction Monitoring in Smart Street Lighting Systems,” in 2018 IEEE 59th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON), 2018, pp. 1–5.

[15] Y. Zhang, R. Su, K. Gao, and Y. Zhang, “Traffic light scheduling for pedestrians and vehicles,” in 2017 IEEE Conference on Control Technology and Applications (CCTA), 2017, vol. 2017-Janua, pp. 1593–1598.

[16] M. Lehtomaki et al., “Object Classification and Recognition From Mobile Laser Scanning Point Clouds in a Road Environment,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 2, pp. 1226–1239, Feb. 2016.

[17] K. Purwanto, I. Iswanto, T. Khristanto, and M. Yusvin, “Microcontroller-based RFID, GSM and GPS for Motorcycle Security System,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 3, pp. 447–451, 2019.

[18] A. N. N. Chamim, M. Heru Gustaman, N. M. Raharja, and I. Iswanto, “Uninterruptable Power Supply based on Switching Regulator and Modified Sine Wave,” Int. J. Electr. Comput. Eng., vol. 7, no. 3, p. 1161, Jun. 2017.

[19] A. N. N. Chamim, D. Ahmadi, and Iswanto, “Atmega16 implementation as indicators of maximum speed,” Int. J. Appl. Eng. Res., vol. 11, no. 15, pp. 8432–8435, 2016.

[20] I. Iswanto, W. S. Agustiningsih, F. Mujaahid, R. Rohmansyah, and A. Budiman, “Accumulator Charging Control with Piezoelectric Based on Fuzzy Algorithm Scheduling,” TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 16, no. 2, p. 635, Apr. 2018.

[21] T. Padang Tunggal, A. Supriyanto, N. M. Zaidatur Rochman, I. Faishal, I. Pambudi, and I. Iswanto, “Pursuit Algorithm for Robot Trash Can Based on Fuzzy-Cell Decomposition,” Int. J. Electr. Comput. Eng., vol. 6, no. 6, p. 2863, Dec. 2016.

[22] A. Kumar, “A Fuzzy based Soft Computing Technique to Predict the Movement of the Price of a Stock,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 319–324, 2018.

[23] K. Iqbal, M. Adnan, S. Abbas, Z. Hasan, and A. Fatima, “Intelligent Transportation System (ITS) for Smart-Cities using Mamdani Fuzzy Inference System,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 94–105, 2018.

[24] Z. Entekhabi and P. Shamsinejadbabki, “FARM: Fuzzy Action Rule Mining,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, pp. 247–252, 2018.

[25] M. A. Ibraigheeth and S. Abdullah, “Fuzzy Logic Driven Expert System for the Assessment of Software Projects Risk,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 2, pp. 153–158, 2019.

Autonomous Monitoring System using Wi-Fi Economic

Michael Ames Ccoa Garay[1]
Electronics and Telecommunication Engineering
Department, Universidad Nacional Tecnologica de Lima
Sur, Lima, Perú

Avid Roman-Gonzalez[2]
Aerospace Sciences and Health Research Laboratory
(INCAS-Lab), Universidad Nacional Tecnologica de Lima
Sur, Lima, Perú

***Abstract*—In this project, it is presented the implementation of an autonomous monitoring system using solar panels and connecting to the network through Wi-Fi. The system will collect meteorological data and transmit in real-time to the web for the visualization and analysis of the results over temperature, humidity, and atmospheric pressure. The system will allow saving time and money, employing decision making and efficiency. For the development of this device, a small platform "Wemos D1" for the internet of things allows easy programming in the platform "Arduino IDE".**

***Keywords*—Wemos d1 mini-skirt; Wi-Fi; sensor; internet**

I. Introduction

An autonomous meteorological station is a sophisticated device that helps to reduce weather uncertainties. This task is accomplished through a data history that helps making decisions by saving time and money.

Worldwide projects exist on meteorological stations autonomous for cultivation fields.

SENCROP is a technology company to cultivation fields in France where it includes collecting data and also shares them through the internet.

Has a lot of characteristics but among which it stands out is the simplicity of the devices.

It is light because it allows moving the stone from one side to another without difficulty.

The data obtained in the fields of cultivation are processed to an algorithm developed by them for the predictions of pests and diseases, giving an import value in the decision making [1].

Local levels there are projects on automatic meteorological stations.

For example, the project of an automatic meteorological station of the "Reserva Biologica Alberto Maberes Brenes" in costa Rica, you will see the value of obtaining data with quality and improvement in the care of a resource of a state [2].

The project of the autonomous monitoring system uses Wi-Fi for collecting meteorological data. The idea is to visualize, analysis, and storage in a simple way, record, and at a low cost, all data. This system can help recognize the type of microclimate that exists in the different areas to know the composition of the place flora and improve the use of vegetable species [3].

The document consists of the implementation of a standing monitoring system using development cards like the "Wemos D1 mini" for the data collections at a low cost.

In the implementation development platform is used as the Wemo D1 mini, a precision sensor, a module of TP4056, solar panel, and a rechargeable lithium battery. The system can measure atmospheric pressure, temperature, and relative humidity. These data are transmitted in real-time to the internet through the connection Wi-Fi to visualize and analyze the meteorological data.

Also, use panels that provide an autonomy of the device.

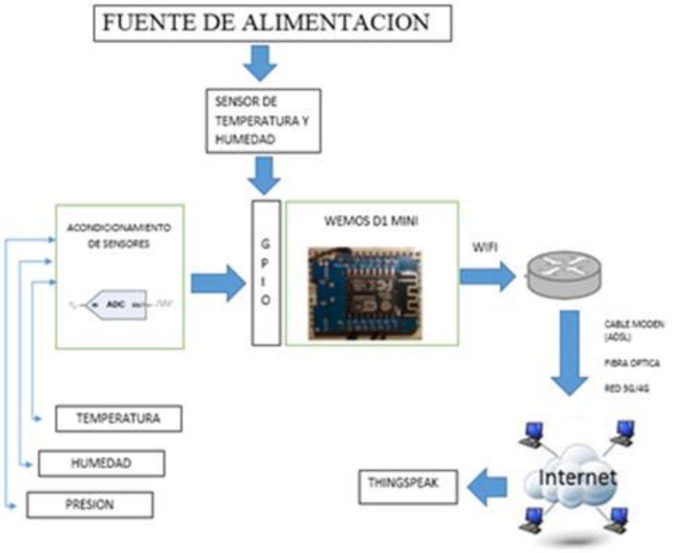


Fig. 1. Diagram Blocks.

II. Methodology

The main components of the system are:

In Fig. 1, a summary of sensor connections is displayed BMP 280, the modules “Wemos” and “TP4056” with the solar panel and the shipping to the internet using the Wi-Fi.

A. Description of main Components

1) Wemos D1 mini-skirt: The Wemos (Fig. 2): Is a development card that incorporates on its printed circuit to ESP-12F. Is small and includes functions useful as a tension regulator of 5V to 3.3V with which the module is fed with 5V. The power source of 5V allows a current of 500mA. All needed consumption is covered, including micro USB and an integrated circuit model CH340G. The CH340G is a converter USB series to connect the plate directly to the computer without the need to through Arduino.

Tension regulator allows feeding directly with 5V without the need of an additional extreme source.

The tension is taken directly of the existence of 5V of the module T4056 that in turn feeds on energy that transfers solar panels with the module “Wemos” connects to the Wi-Fi network and take measures through the sensor [4]. All these features can be seen in Table I.

TABLE. I. Features of the Module Wemos D1 Mini

Parameter	**Value**
Voltage of nutrition	**5V AD**
Voltage of prominent entrance	3.3V AD
Digital pins GPIO	11
Analogical pins ADC	1
Average of consumed current	70 mA
Memory flash external	4MB
Date RAM	96kB
I mince current	400mA
Frequency of clock	80 Mhz 160 Mhz
Microcontroller	ESP8266
Dimensions	34.2mm x 25,6 mm
I weigh	10g

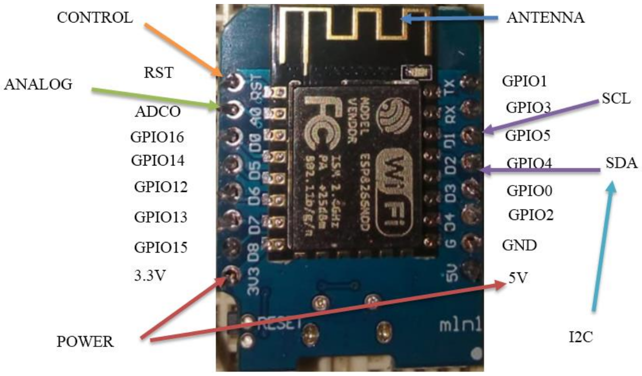


Fig. 2. Module Wemos D1 Mini.

2) I modulate TP4056: The modulate is a battery charger of lithium TP4056 that account with an entry micro USB and an additional two contactors for direct connection IN+ and IN-. At the other end are the terminals B+ and B- that is responsible for connecting to the battery to change. The modulate loads the constant current 1A until the moment the current starts to decrease, activating the mode to change at a constant voltage. All these features can be seen in Table II and in Fig. 3.

3) Sensor BMP280: The sensor BMP 280 will allow us measuring atmospheric pressure in the status of 300 to 1100hpa with +/- average error 40 to 85 C. With the error one can measure 1hpa regarding temperature from -1,0 C. The interface allows us connecting the Wemos by SPI or I2C that needs a source of 3.3V [5]. All these features can be seen in Table III and in Fig. 4.

4) Power supply: For this project, it was used a battery of 2800 mah. That connect through the module TP40556 will get energy from solar panels. It can be seen in Fig. 5.

TABLE. II. Characteristics of the I Modulate TP4056

Parameter	**Value**
Voltage from the start	4.5V 5.5V
Charging voltage full	4.2V
Entrance Microcomputer Usb	If
Operating temperature	- 10 C to +85 C
Inverse polarity	No
Charging precision	1,5 %
Charging mode	Linear load
Current of maximum charging exit	1.2A

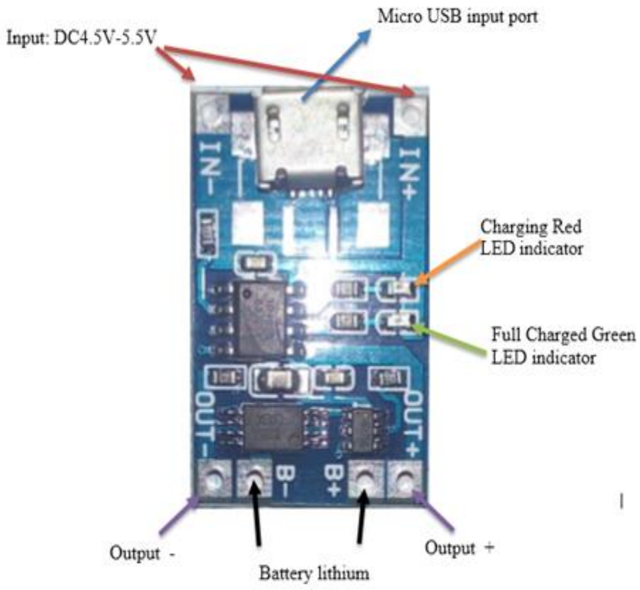


Fig. 3. Module TP4056.

TABLE. III. CHARACTERISTICS OF THE SENSOR BMP 280

Parameter	**Value**
Voltage of operation	1.8V 3.3V AD
Interface of communication	I2C or SPI 3.3V
Status of pressure	300 to 1100hPa
Absolute precision	1 hPa
Measurement of temperature	- 40 C to +85 C
Precision of temperature	1 C
Frequency of sampling	157 Hz
I decrease consumption of energy	If

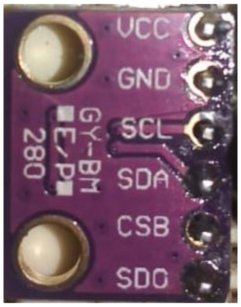


Fig. 4. Sensor BMP280.

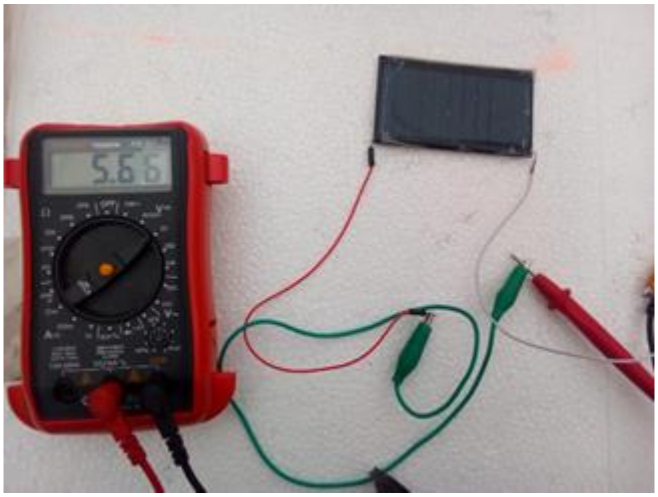


Fig. 5. Solar Panel of 5.5V.

B. *Implementation of the Autonomous Monitoring System*

The monitoring system is autonomous; there must be a continuous power supply. The best way to provide uninterrupted power to the circuit is through the use of a battery. After a few days, the battery charge would be depleted, and it is very complicated to get the energy of the sun to charge the batteries and provide power to all circuit. For this project, it was used a battery of lithium 2800 mah. It can be seen in Fig. 6.

The system can measure temperature, humidity, barometric, pressure. Also, it can monitor the meteorological parameters due to a record stored from the web.

The battery is charged using a solar panel through a charging module TP4056.

Fig. 6. Lithium Battery of 2800mA.

The module TP4056 is ideal for loading cells (Solar panels) from 3.7V and 1A. This module will offer a constant charging current of 1A and then cut when the load is also finished. When the tension of the battery descends below 2.4V, the IC of protection will reduce the load to protect her cells of the battery against the low voltage. Also, protect against the excessive energy and the connection of inverse polarity. It can be seen in Fig. 7.

1) Installation of the solar panel and the battery: A cable is welded to the solar panel's negative terminal to the positive terminal and the black wire. Next, insert the battery's support in the slot in the protoboard's part. It can be seen in Fig. 8.

2) Programming: After the implementation of the plate of development Wemos, the sensor should position itself. Then, it will send the instructions for the sensor of ultrasounds to accomplish the measurement of temperature, humidity, and atmospheric pressure for which washed out to stub the following code. The Wemos module gets connected to net road Wi-Fi and besides sends the data for its visualization in real-time. The inserted code is shown from Fig. 10 to Fig. 15.

To use Wemos D1 with the library, Arduino, you will have to use the ÍDE ARDUINO with support ESP8266. Other forms of plate seriously install the plate's support ESP8266 in the ÍDE of Arduino [6]. It can be seen in Fig. 9.

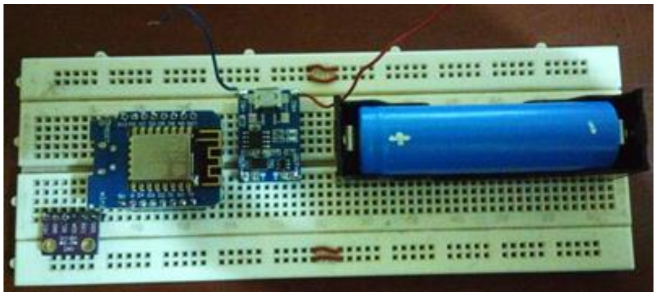


Fig. 7. Prototype Implementad unrelated.

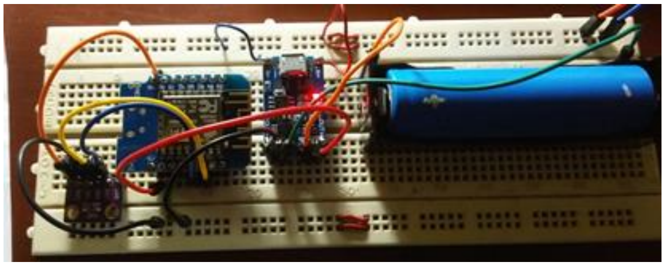


Fig. 8. Prototype Implemented without Connections.

The following adjustments are preferable:

- The PU's frequency:

  80MHz 160MHz

- Size of the flash: The archival system's (3M SPIFFS) - the archival system's Size 3M 4M (1M SPIFFS) - So Big a 4M 1M

- Charging velocity: 921600 bps [7]

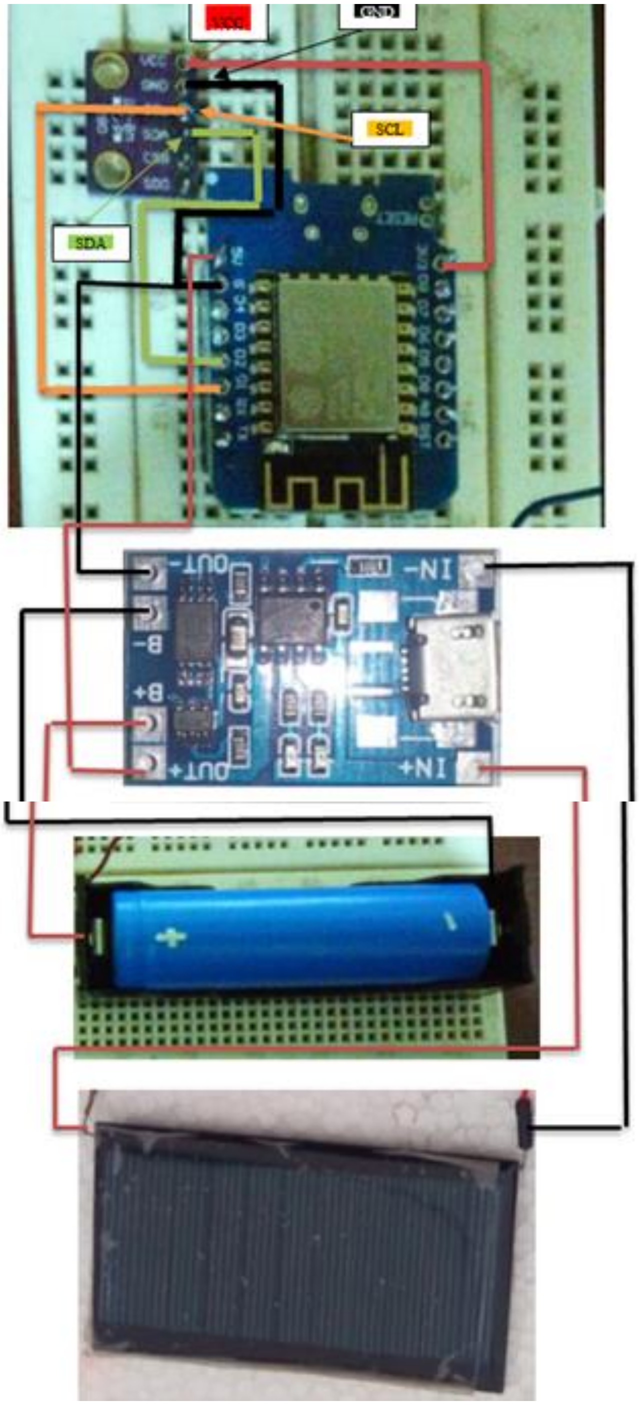
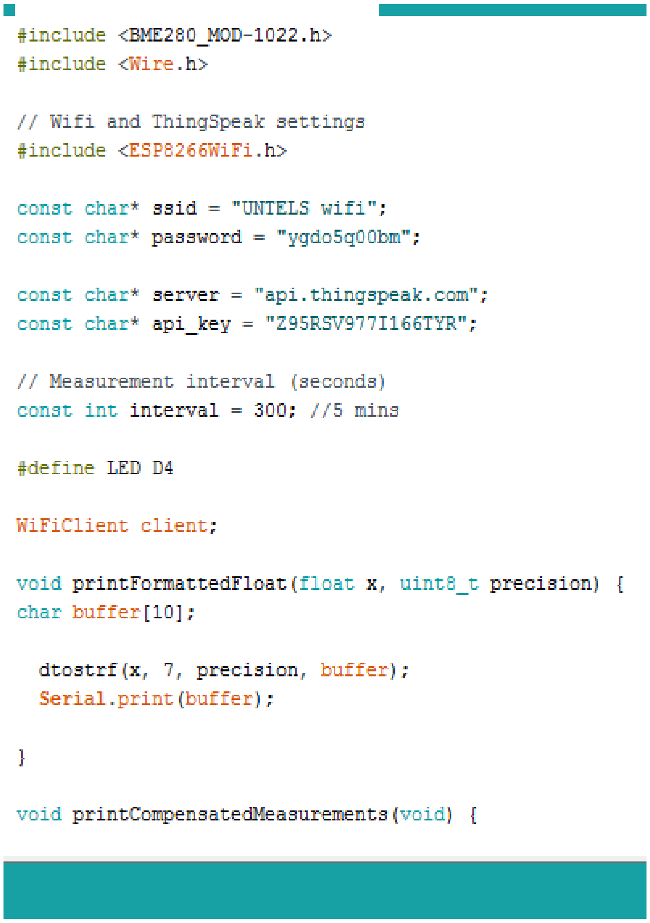
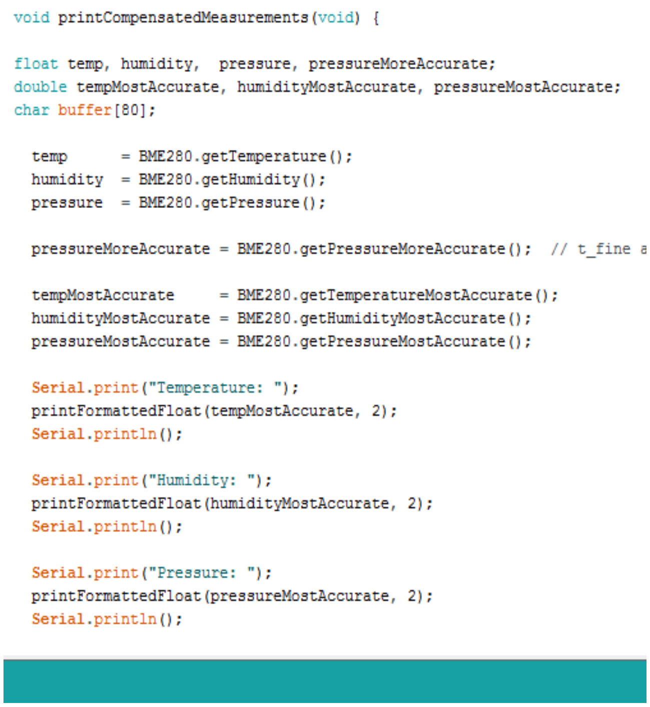


Fig. 9. Wire Diagram.



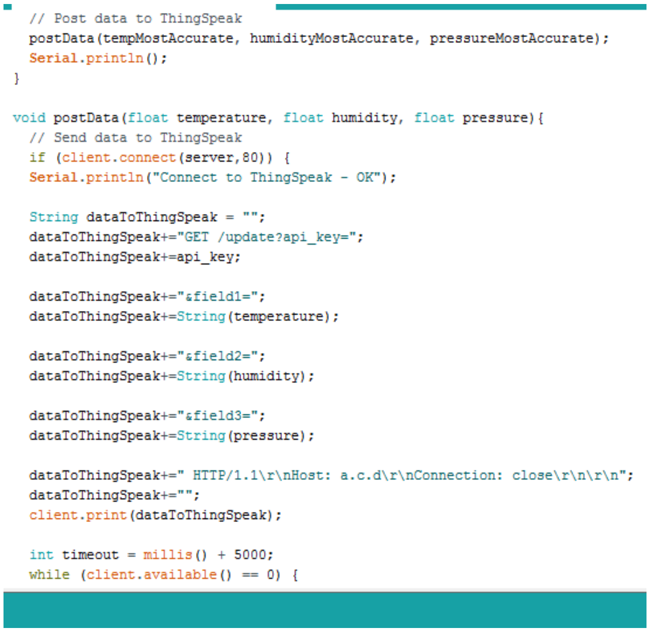
```
#include <BME280_MOD-1022.h>
#include <Wire.h>

// Wifi and ThingSpeak settings
#include <ESP8266WiFi.h>

const char* ssid = "UNTELS wifi";
const char* password = "ygdo5q00bm";

const char* server = "api.thingspeak.com";
const char* api_key = "Z95RSV977I166TYR";

// Measurement interval (seconds)
const int interval = 300; //5 mins

#define LED D4

WiFiClient client;

void printFormattedFloat(float x, uint8_t precision) {
char buffer[10];

  dtostrf(x, 7, precision, buffer);
  Serial.print(buffer);

}

void printCompensatedMeasurements(void) {
```

Fig. 10. Network Configuration that One is Going to Connect.



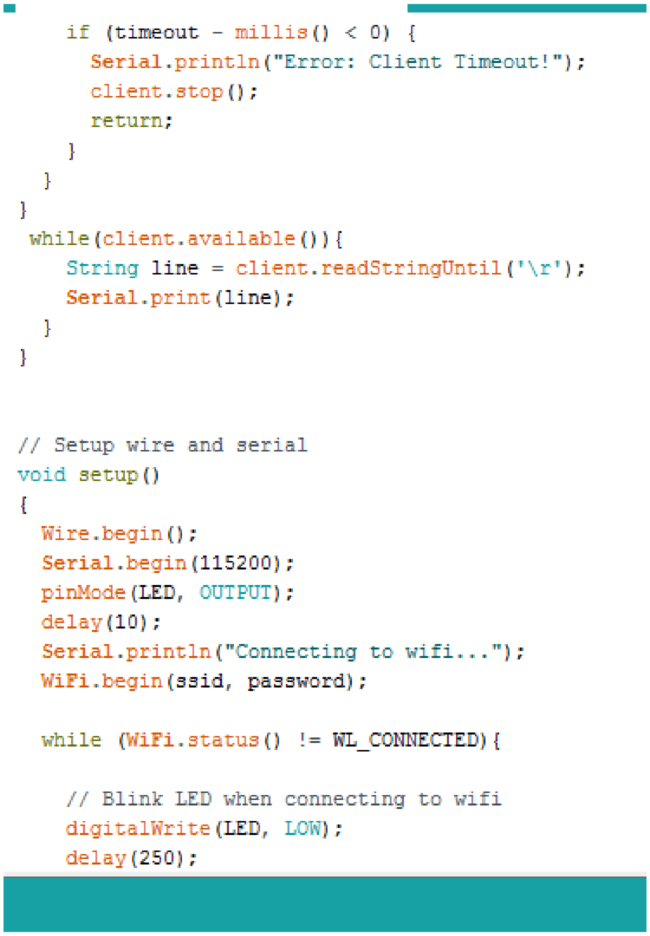
```
void printCompensatedMeasurements(void) {

float temp, humidity,  pressure, pressureMoreAccurate;
double tempMostAccurate, humidityMostAccurate, pressureMostAccurate;
char buffer[80];

  temp      = BME280.getTemperature();
  humidity  = BME280.getHumidity();
  pressure  = BME280.getPressure();

  pressureMoreAccurate = BME280.getPressureMoreAccurate();  // t_fine a

  tempMostAccurate     = BME280.getTemperatureMostAccurate();
  humidityMostAccurate = BME280.getHumidityMostAccurate();
  pressureMostAccurate = BME280.getPressureMostAccurate();

  Serial.print("Temperature: ");
  printFormattedFloat(tempMostAccurate, 2);
  Serial.println();

  Serial.print("Humidity: ");
  printFormattedFloat(humidityMostAccurate, 2);
  Serial.println();

  Serial.print("Pressure: ");
  printFormattedFloat(pressureMostAccurate, 2);
  Serial.println();
```

Fig. 11. Declaring Variables.



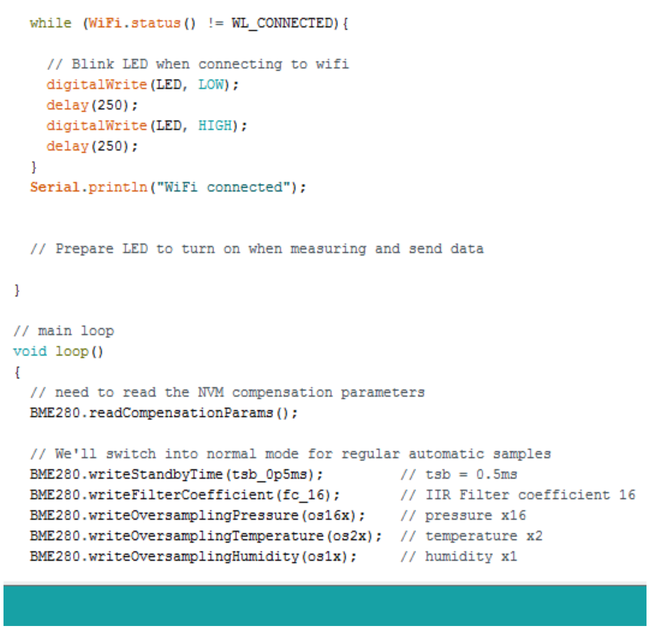
```
  // Post data to ThingSpeak
  postData(tempMostAccurate, humidityMostAccurate, pressureMostAccurate);
  Serial.println();
 }

void postData(float temperature, float humidity, float pressure){
  // Send data to ThingSpeak
  if (client.connect(server,80)) {
  Serial.println("Connect to ThingSpeak - OK");

  String dataToThingSpeak = "";
  dataToThingSpeak+="GET /update?api_key=";
  dataToThingSpeak+=api_key;

  dataToThingSpeak+="&field1=";
  dataToThingSpeak+=String(temperature);

  dataToThingSpeak+="&field2=";
  dataToThingSpeak+=String(humidity);

  dataToThingSpeak+="&field3=";
  dataToThingSpeak+=String(pressure);

  dataToThingSpeak+=" HTTP/1.1\r\nHost: a.c.d\r\nConnection: close\r\n\r\n";
  dataToThingSpeak+="";
  client.print(dataToThingSpeak);

  int timeout = millis() + 5000;
  while (client.available() == 0) {
```

Fig. 12. A Code is Generated for the Website.



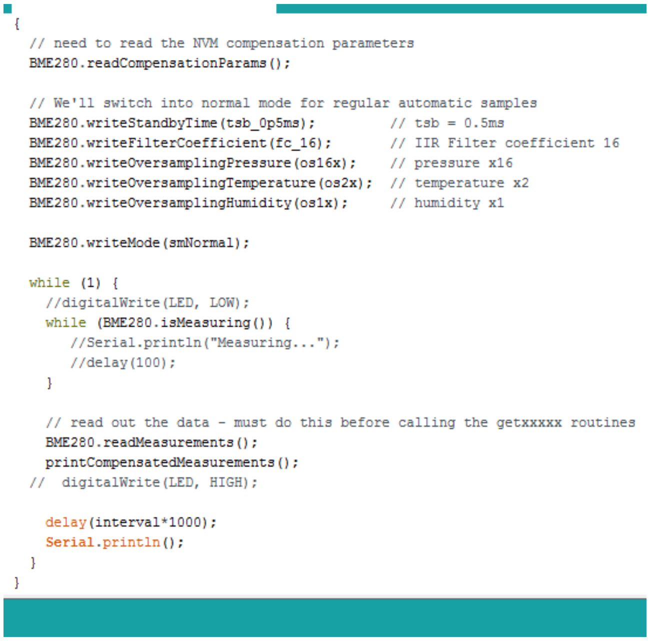
```
    if (timeout - millis() < 0) {
      Serial.println("Error: Client Timeout!");
      client.stop();
      return;
    }
  }
}
 while(client.available()){
    String line = client.readStringUntil('\r');
    Serial.print(line);
  }
}


// Setup wire and serial
void setup()
{
  Wire.begin();
  Serial.begin(115200);
  pinMode(LED, OUTPUT);
  delay(10);
  Serial.println("Connecting to wifi...");
  WiFi.begin(ssid, password);

  while (WiFi.status() != WL_CONNECTED){

    // Blink LED when connecting to wifi
    digitalWrite(LED, LOW);
    delay(250);
```

Fig. 13. Showing Wi-Fi Connection.



```
  while (WiFi.status() != WL_CONNECTED){

    // Blink LED when connecting to wifi
    digitalWrite(LED, LOW);
    delay(250);
    digitalWrite(LED, HIGH);
    delay(250);
  }
  Serial.println("WiFi connected");


  // Prepare LED to turn on when measuring and send data

}

// main loop
void loop()
{
  // need to read the NVM compensation parameters
  BME280.readCompensationParams();

  // We'll switch into normal mode for regular automatic samples
  BME280.writeStandbyTime(tsb_0p5ms);         // tsb = 0.5ms
  BME280.writeFilterCoefficient(fc_16);       // IIR Filter coefficient 16
  BME280.writeOversamplingPressure(os16x);    // pressure x16
  BME280.writeOversamplingTemperature(os2x);  // temperature x2
  BME280.writeOversamplingHumidity(os1x);     // humidity x1
```

Fig. 14. Showing Network Connection.



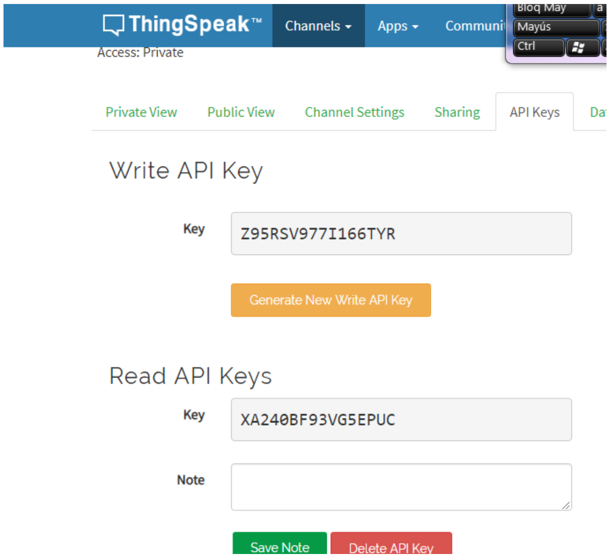
```
{
  // need to read the NVM compensation parameters
  BME280.readCompensationParams();

  // We'll switch into normal mode for regular automatic samples
  BME280.writeStandbyTime(tsb_0p5ms);         // tsb = 0.5ms
  BME280.writeFilterCoefficient(fc_16);       // IIR Filter coefficient 16
  BME280.writeOversamplingPressure(os16x);    // pressure x16
  BME280.writeOversamplingTemperature(os2x);  // temperature x2
  BME280.writeOversamplingHumidity(os1x);     // humidity x1

  BME280.writeMode(smNormal);

  while (1) {
    //digitalWrite(LED, LOW);
    while (BME280.isMeasuring()) {
       //Serial.println("Measuring...");
       //delay(100);
    }

    // read out the data - must do this before calling the getxxxxx routines
    BME280.readMeasurements();
    printCompensatedMeasurements();
  //  digitalWrite(LED, HIGH);

    delay(interval*1000);
    Serial.println();
  }
}
```

Fig. 15. Constant Repetition with the Command.

3) Sending the data of the sensor BMP 280 to the Web Thingspeak

First, an account in Thingspeak is created. Next, a new canal in the account of Thingspeak is created. It gets stung with the data at the

- Field 1: Temperature
- Field 2: Humidity
- Field 3: Pressure



Write API Key

Key: Z95RSV977I166TYR

Read API Keys

Key: XA240BF93VG5EPUC

Note:

Fig. 16. Nail Down API.

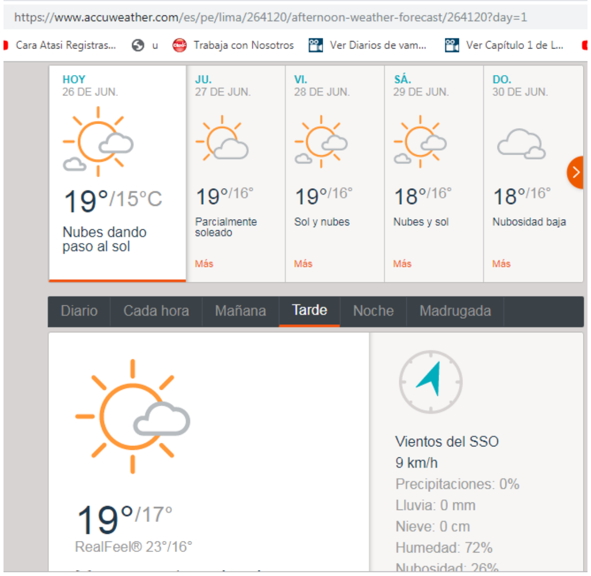
The channel is selected after in Thingspeak's new account. It stops next looking for eyelash password API and imitates the keyboard write.

The net which one wants to connect itself the SSID to is written in the interface of the Arduino Íde once the code was opened stops next placing the password on the code. API replaces the WRITE itself and API copies the write itself key that Thingspeak's page provides us. It is essential to have installed the BMP's bookstores 280 [8]. It can be seen in Fig. 16.

III. Results

ThingSpeak compiles information about temperature, humidity, and atmospheric pressure himself next after being sent to the platform the data in graphs could be visualized.

Data of temperature he shows the following values shown in Fig. 17 and Fig. 18.



HOY 26 DE JUN. 19°/15°C Nubes dando paso al sol

JU. 27 DE JUN. 19°/16° Parcialmente soleado

VI. 28 DE JUN. 19°/16° Sol y nubes

SÁ. 29 DE JUN. 18°/16° Nubes y sol

DO. 30 DE JUN. 18°/16° Nubosidad baja

19°/17°
RealFeel® 23°/16°

Vientos del SSO
9 km/h
Precipitaciones: 0%
Lluvia: 0 mm
Nieve: 0 cm
Humedad: 72%
Nubosidad: 26%

Fig. 17. The Platform's Graph AccuWeather.

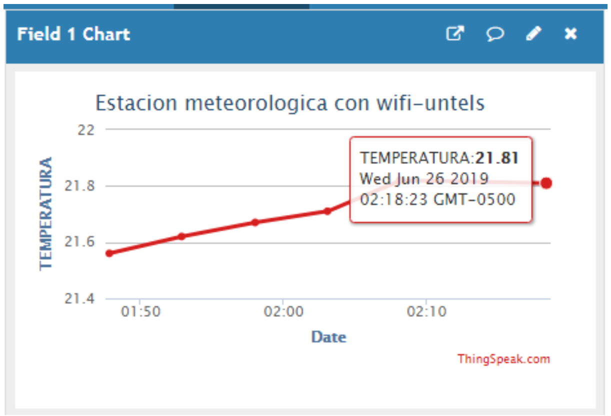


Fig. 18. Graph ThingSpeak of Temperature.

IV. Discussion

The system's implementation is simple and economical. The objective to gather meteorological data and to be able to visualize, utilizing connection Wi-Fi to the internet, from any place thanks to the card of development fulfills Wemos. Wemos has the individual capacity of no. to control to transfer the data to the net giving one a great variety of application software. The sensor BMP280 is the precise sensor that if not one, also obtains the meteorological measures.

The application software that one can get from the module Wemos was joined of ThingSpeak. It is one of the platforms of software enlarged for the ones that wants to start-up in the world of the internet of thing (IoT). It is a simple way due to his compatibility with the card Arduino.

The recommendation is to take into account the bringing up to date of the bookstores of the Arduino. It would have the ones one come than by default that to make some modifications, but it is easy to obtain the necessary bookstores in the web.

The system has wholes the elements to be able to improve from their web interface in the designing improvement as in the structure and components. They allow protecting of the storms of the climate once the external use was given to and your portability.

Data of humidity shows the following values shown in Fig. 19.

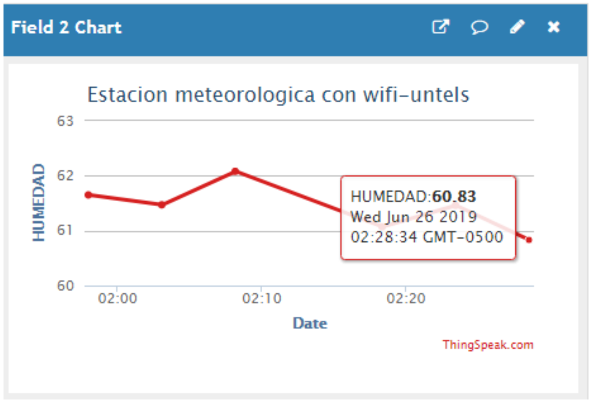


Fig. 19. Graph ThingSpeak of Relative Humidity.

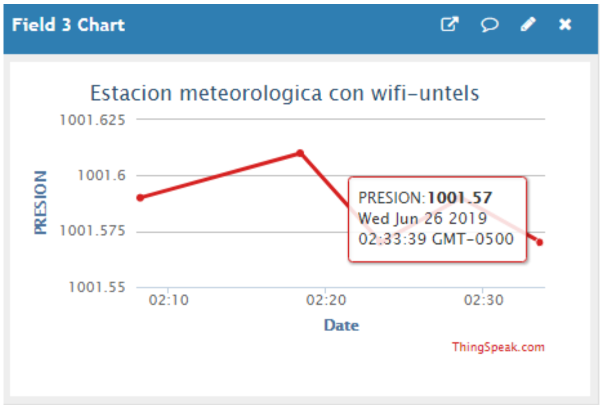


Fig. 20. Graph ThingSpeak of Atmospheric Pressure.

Data of atmospheric pressure shows values shown in Fig. 20.

V. Conclusions

ThingSpeak compiles information about temperature, humidity, and atmospheric pressure. This itself is next after being sent to the platform and could compare with the data in AccuWeather, a North American company, and renders commercial services of weather forecast all over the world. One can become evident in the image than the sensor Bmp 280 is within range regarding the company of meteorological. They are strong values that the project gives us but unlike the rest of meteorological devices. In general, a technical service that one can help when a flaw exists does not have. This system is a very cost-reducing and easy project of implementation; besides that it is straightforward since one can move without worrying about the reserve of energy. One can pick up data that are very important for the study of farm cultivation for deeply letting us know the acquaintance microclimate that the Peruvian ground has.

References

[1] Tripathi, Harish Kumar, and B. K. Sen. "Crop science literature and Bradford law." Annals of Library and Information Studies (ALIS) 63.2 (2016): 85-90.

[2] Sandoval, Chapoñan, and Julio Cesar. "Caracterización temporal del viento registrado en el borde costero de la ciudad de Santa Rosa y en las Islas Lobos de Afuera, durante los años 2005 al 2012." (2016).

[3] Sala, Juan Ignacio, et al. "Estación de medición para análisis y control de parámetros ambientales." IX Congreso Argentino de AgroInformática (CAI 2017)-JAIIO 46-CLEI 43 (Córdoba, 2017). 2017.

[4] Girón, Luis Diego Maldonado. "Diseño De Una Estación Meteorológica Con Control De Accionamientos Electromecánicos Y Monitoreo De Sensores."

[5] Medina Marín, Jairo José, and Alexis Oldemar Valle Ruiz. Diseño de una estación meteorológica para la medición de temperatura, humedad, presión atmosférica y radiación solar en el área del Recinto Universitario Rubén Darío, la cual estaría ubicado en el Centro de Investigaciones Geo Científicas (CIGEO), en la UNAN Managua. Diss. Universidad Nacional Autónoma de Nicaragua, Managua, 2017.

[6] Claudio P.Millahual,"Arduino-De cero a Experto:Proyectos Practicos-electronica , hardware y programacion",pp-110-114.

[7] Ben Akka, Youssef, et al. "Control and Command Of Several Greenhouses Via Telegram Messenger." Proceedings of the 2nd International Conference on Networking, Information Systems & Security. ACM, 2019.

[8] Loureiro Garrido, Rubén. "Estudio Plataformas IoT."

Impact of ICT on Students’ Academic Performance: Applying Association Rule Mining and Structured Equation Modeling

Mohammad Aman Ullah[1], Mohammad Manjur Alam[2], Ahmed Shan-A-Alahi[3]
Mohammed Mahmudur Rahman[4], Abdul Kadar Muhammad Masum[5], Nasrin Akter[6]
Dept. of Computer Science and Engineering
International Islamic University Chittagong
Chittagong-4203, Bangladesh

***Abstract*—Information and communication technology (ICT) plays a significant role in university students’ academic performance. This research examined the effect of ICT on the students’ academic performance at different private universities in Chittagong, Bangladesh. Primary data have been collected from the students of those universities using a survey questionnaire. Descriptive Statistics, Reliability Analysis, Confirmatory Factor Analysis, OLS regression, Structured Equation Modeling (SEM) and Data Mining algorithms such as Association rule mining and éclat have been employed to evaluate the comparative importance of the factors in identifying the academic performance of the students. From a statistical and mining perspective, overall results indicate that there is a significant relationship between ICT use and students’ academic performance. Also, student’s addiction to ICT has a significant influence on the comparative measurement in identifying the academic performance of the students. Finally, some recommendations are provided on the basis of the findings.**

Keywords*—*Information and Communication Technology (ICT); student; academic; performance; association rule mining

I. Introduction

The application and effect of Information and Communication Technology (ICT) is considered to be a topic of interest in different areas of real life mostly in education. Educators can now use ICT as a tool that allows modifying the instructional approach in the classroom in order to get better students’ performance. Learning institutions are adopting ICT based instructional approach and presenting ICT oriented academic programs. Recently, the Government of Bangladesh has also adopted the use of ICT in the Bangladeshi educational Institutions (home and abroad), giving the importance on this topic. Therefore, students own the ICT facilities for both the academic and non-academic purposes using diverse smart devices and the internet. The use of the ICT in both the academic and non-academic purposes poses the demands to evaluate the students’ honesty and the academic performance. Therefore, the objectives of this research are:

1) To uncover the effect of ICT on the academic performance of students at various universities in Chittagong, Bangladesh.

2) To find the relationship between the use of ICT for academic and non- academic purposes by students

To meet the objectives, the data were collected from different private universities of Chittagong, Bangladesh. Descriptive statistics, reliability analysis, Confirmatory Factor Analysis, OLS regression, path analysis and data mining algorithms such as Association rule mining and éclat have been employed to weigh up the comparative significance of the features. The overall results from statistical and data mining analysis indicated that, there is a statistically significant relationship between ICT use and students’ academic performance. Also, student’s addiction to ICT has a significant influence in comparative measurement in identifying the academic performance of the students. But, the overuse of ICT hampers outcome to a large extend.

II. Background

The inadequate or lack of ICT facilities appears as significant barriers in students ICT usage. Although students’ perceptions are explained in diverse studies as important significant variables to analyze ICT usage, which only depends on enough ICT facilities [7-10], [17], [27]. In conducting multi-media classrooms ICT infrastructure was found to play a significant role [17]. However, Students are not using ICT always for an academic purpose; however, it can be used for the different purposes. For instance, students might use ICT to make class equipment or for individual use [27]. The application of ICT and its effect on student throughput in university education was not reflected to be ideal so far; rather it shows mixed results in the previous study. Previous research has failed to give a logical idea regarding the con- sequence of ICT on students’ success. Firstly, some literature could not show a real outcome of ICT on students’ perfor- mance in university education. There are very rare experiential support about the influence and efficacy of ICTs on students’ academic performance at university level both in developed and underdeveloped countries [29-31]. The research done so far does not clearly reflect pure effects of ICT on student’s academic achievements [25], which on the other hand demands the synchronization and regularity regarding the ICTs effect due to methodological restrictions [3]. There are many factors found by different empirical studies for performance improvement at the university level, but neither of the study emphasizes on the ICT use, rather emphasized on how it was used [23]. The research related to the use of ICT in teaching and learning by teachers has been done in several surveys [6] [26].

As per the surveys conducted by [24][12], the students of colleges and universities in developed countries use the ICT in all their learning activities. But, due to the over accesses of amusing materials through ICT tools hamper their overall performance [14][16][19][28]. They have conducted their survey in Bangladesh and Indonesia, respectively, and found that, about 80% of the respondents do not agree that the ICT is useful in improving the academic performance; they rather think it as a source of entertainment. In [1], the authors reported the totally different result. According to his survey of various Bahraini universities, students are more motivated to learn through ICT and thus improve students' performance. A Similar result was found by [11][13][22]. A Result Prediction System was developed by [4] to carry out association rule mining automatically on the collection of earlier student's results and predicts the current students' results. Before it, they have also clustered the subjects on the basis of unique criteria of the subjects. A study was conducted on 320 undergraduate students in Ghana to find the impact of some selected ICT devices in students' academic performance. They have conducted statistical experiments such as descriptive statistics and regression in their study and found that, tools such as email intensify the student's academic performance [18].

Consequence of ICT on students' academic outcome of four Saudi Universities was investigated by [5] in the study they have used structure equation modeling for validating their research model. Their findings show that use of ICT increases students' performance, in particular women, but the university IT course has no influence on the overall academic outcome.[2] applied the propensity score matching method to identify the consequence of ICT use on academic achievement by the school students in Argentina. Their study found no significant relationship between ICT use and academic program. So, it is evident from the literature review that, the performance may or may not improve due to ICT use; it also depends on other factors. This study tries to find the exact association between these factors and impart the real picture.

III. Methodology and Data Collections

The data have been collected for this study from different private universities of Chittagong, Bangladesh such as International Islamic University Chittagong (39.5%), Premier University Chittagong (21.4%), BGC Trust University (18.6%), Port City International University (9%), Chittagong Independent University (9%), and East Delta University (2.4%,) from 2nd to 8th semester students. A total of 210 student data has been collected randomly using a structured questionnaire during the spring 2018 semester. For determining the sample size, the "rule of thumb is larger than 30 and less than 500" [20-21] was maintained. Questionnaires were designed by two parts. Part A was the basic information and part B was the different dimension of students' academic performance by using the 5-point Likert scale. Part A was collected using the features such as Name, University Name, Semester, Department, Gender, Use of Technology, Regularly Internet Browsing Time, Browsing Period, and CGPA. On the other hand, the information used in part B has been discussed in Table II. The study was performed using the Statistical Package for Social Sciences (IBM SPSS Statistics, Version 24.0), STAT 12, Smart PLS3 and R programming language.

IV. Result and Analysis

A. Background Characteristics of Respondents

This is undertaken with a view to giving an idea about the dataset. Percentage method was used to describe the background characteristics of this dataset. The implication of Table I is that most of the respondents 132 (62.9%) were males while only 78 (37.1%) of them were females. The study also implies that, 26.7% of students use the internet for about 1 to 2 hours, 24.8% for three hours, and 47.6% for more than 4 hours daily. Table I also reveal that 96.7% students browse the internet regularly for different purposes. According to the findings of the research, half (50.5%) of the students uses a laptop on their academic purpose, 34.3% have access only to the internet for their academic purpose, 49.0% of the students use the internet solely for academic purposes. 67.7% of the students' use the ICT in non-academic purpose. This research found that, students with CGPA <3 is 28.1%, 3-3.3 is 38.6% and 3.5- 4.00 is 33.6%.

B. Descriptive Statistics

The individual items in the Questionnaire, Indicators of five original dimensions and their means, standard deviations, and result of reliability items (Cronbach's Alpha if Item deleted) are performed in Table II. The overall mean and standard deviation of different items is 1.946 and 0.626 respectively. The Result of reliability items in SPSS shows that, there is an internal consistency between the items in questionnaire related to the students' academic performance. The overall Cronbach's Alpha is 0.844.

TABLE. I. Basic Information Profile

Variables	**Category of variable**	**Frequency**	**Percent (%)**
Gender	Male	132	62.9
	Female	78	37.1
Internet Browse Regularly	Yes	203	96.7
	No	7	3.3
Browsing Period(Daily)	Zero hour	2	1
	1 to 2 hours	56	26.7
	Three hours	52	24.8
	Four or more	100	47.6
Use of Technology	Laptop	106	50.5
	Desktop	30	14.3
	personal mobile	72	34.3
Average CGPA	Less than 3	59	28.1
	3.0-3.5	81	38.6
	>3.5	70	33.3
Internet use in Non-Academic Purpose	Yes	142	67.7
	No	68	32.3

TABLE. II. DESCRIPTIVESTATISTICS OF INDIVIDUAL ITEM SCALE

Individual Items in Questionnaire	Item indicators	Mean	Standard Deviation	Cronbach's Alpha if Item Deleted
ICT in Class Room	IC	3.219	1.003	0.827
Multimedia classroom	IC1	3.362	1.146	0.828
To count attendance in a class	IC2	3.267	1.285	0.825
To achieve the curriculum goals	IC3	3.286	1.273	0.826
Use of ICT by teacher during lecture	IC4	3.286	1.266	0.831
To take class by the projector	IC5	2.895	1.057	0.827
Outside the Class Room	OC	2.268	0.611	0.836
To prepare homework or assignment	OC1	3.21	1.235	0.843
Preparation of examination	OC2	2.762	1.054	0.834
Watching academic lecture	OC3	2.605	1.008	0.834
Field study	OC4	2.762	1.007	0.834
Opinion Towards ICT Use	OT	1.114	0.54	0.836
I spent most of the time with ICT	OT1	2.719	2.183	0.835
Share personal presentation and information.	OT2	2.852	1.425	0.837
Addiction of ICT	Ad	1.215	0.451	0.837
Using ICT my academic productivity decreased	Ad1	3.224	1.288	0.843
Playing game in online	Ad2	2.852	1.328	0.83
Academic Impacts	AI	1.913	0.523	0.847
ICT improves students' performance on examinations	AI1	3.176	1.317	0.85
Using ICT subjective knowledge is good	AI2	3.144	1.278	0.842
To find scholarship	AI3	3.276	1.206	0.849
Overall	210	1.946	0.626	0.844

Dependent variable: Academic Impact

C. *Confirmatory Factor Analysis (CFA) and Structured Equation Modeling (SEM)*

CFA has been used to justify the model fit of the hypothesized structure. The following hypothesis has been considered in the Structural equation model:

Hypothesis: There is no statistically significant relationship between the students' use of ICT and their academic performance.

The model shown in Fig. 1 describes the internal consistency of different items along with regression coefficients and R-square value. In order to evaluate the internal consistency of the factor loading, most of the factors exceed the limiting value 0.70 [32] are shown except IC5, OC1, OC2, OT1 AI1, and AI3. The factors IC5, OC1, OC2, OT1, AI1, and AI3 are considered, because their result of reliability items on Cronbach's Alpha is greater than 0.70. The proposed model explained 34.5% variance in the academic impact of ICT. Under the reflective measurement model, cronbach's alpha, composite reliability, and the average variance extracted (AVE) are assessed. From Table III it is shown that although Cronbach's Alpha of one variable is low, but composite reliability and the AVE satisfy the minimum cutoff value 0.7 and greater than 0.5 [32]. Although reliability values greater than 0.70 is good, but between 0.60–0.70 is also acceptable if another dimension of the construct's validity is good [15].

Table IV indicates that all construct shows satisfactory discriminant validity, where the diagonal value is larger than the correlations (off-diagonal) for all reflective constructs [32]. Table V reports that, multi-collinearity has not been detected for independent variables (Addiction of ICT, ICT in Class Room, Opinion towards ICT Use, Outside the Class Room). Table VI shows the overall fit of the structural model. It is evident that, addiction of ICT is the most influential effect on students academic impact β = 0.565, followed by Opinion towards ICT use in academic purpose β = 0.116. But ICT use in inner and outer in the classroom has not been significant. The probable reason is that, the teachers' do not use ICT tools in the classrooms or the ICT facilities in different institutions are not well enough. The overall structural model has shown in Fig. 2.

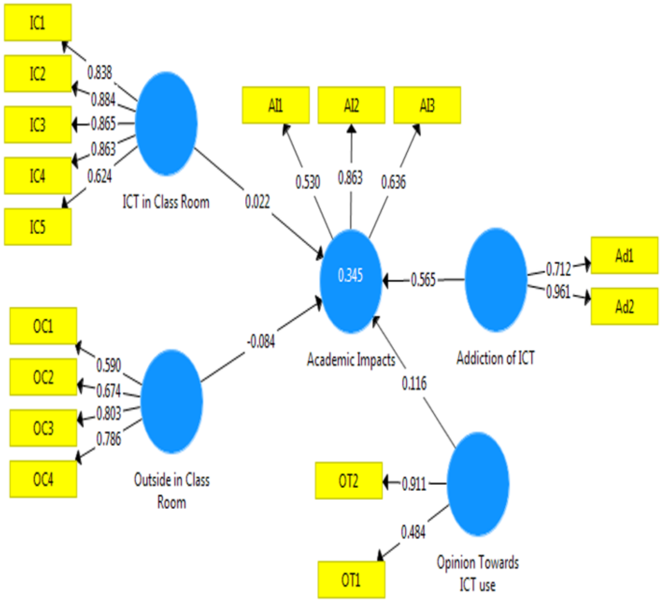


Fig. 1. The Hypothesized Structured Model.

TABLE. III. CONSTRUCT RELIABILITY AND VALIDITY

Loading Variables	Cronbach's Alpha	Composite Reliability	Average Variance Extracted (AVE)
Academic Impacts	0.676	0.724	0.807
Addiction of ICT	0.837	0.831	0.715
ICT in Class Room	0.879	0.91	0.673
Opinion Towards ICT Use	0.697	0.775	0.632
Outside the Class Room	0.713	0.808	0.616

TABLE. IV. FACTOR MATRIX SHOWING DISCRIMINANT VALIDITY

Variables	Academic Impacts	Addiction of ICT	ICT in Class Room	Opinion Towards ICT Use	Outside the Class Room
Academic Impacts	0.691				
Addiction of ICT	0.578	0.846			
ICT in Class Room	0.228	0.372	0.821		
Opinion Towards ICT Use	0.252	0.301	0.367	0.729	
Outside the Class Room	0.187	0.355	0.555	0.504	0.718

TABLE. V. COLLINEARITY STATISTICS (VIF)

Variables	VIF
Addiction of ICT	1.225
ICT in Class Room	1.535
Opinion Towards ICT Use	1.384
Outside the Class Room	1.731

TABLE. VI. PATH COEFFICIENT

Relationship	Coefficient	Standard Error	T Statistics	P Values
ICT in Class Room -> Academic Impacts	7.5	0.059	0.366	0.357
Outside the Class Room -> Academic Impacts	-0.084	0.07	1.198	0.116
Opinion Towards ICT Use -> Academic Impacts	0.116	0.069	1.692	0.043
Addiction of ICT -> Academic Impacts	0.565	0.053	10.648	0

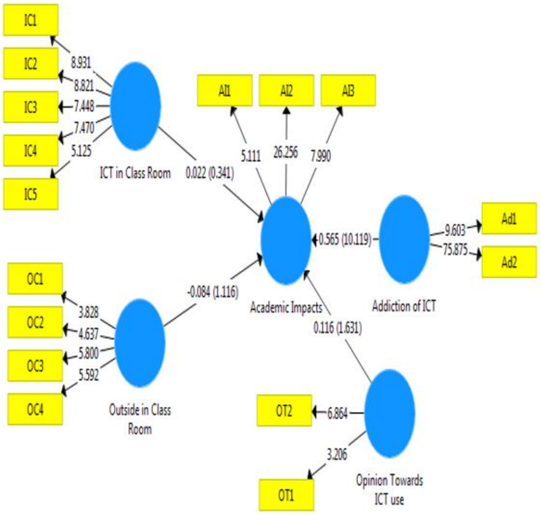


Fig. 2. Path Coefficient with T-Value.

D. Students Performance Estimation using OLS Regression

Ordinary least squares regression analyses were also carried out in this study. The academic impact was used as dependent variable and addiction of ICT, ICT in the classroom, opinion towards ICT use, and ICT use outside the classroom was used as independent variables. The proposed model is

$$Y = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 +U \quad (1)$$

Where, the dependent variable Y is the Academic impact of ICT, X1 = ICT in the classroom, X_2 = ICT use outside the classroom, X_3 = Opinion towards ICT Use, X_4 = Addiction of ICT, and U= Disturbance term. The details' regression results have been given below.

Table VII shows the F test for the regression model which evaluates the statistical significance of the overall regression model. The F-value is the mean square regression (3.483) divided by the Mean Square Residual (0.211). The p-value associated with this F value is very small (0.000). The value of R-square is 0.243 reflects that 24% of the variation in academic impacts can be predicted from the combination of independent variables ICT in the classroom, outside the classroom, opinion towards ICT use, addiction of ICT. Table VIII shows that, the relative importance of significant dimension is observed by the regression coefficient. By analyzing the results, it is shown that, the variable Addiction of ICT has the highest regression coefficient with the significance p- value (p 0.01). This means that "Addiction of ICT" dimension is the main predictor in the estimation of OLS regression. All the other variables are not significant.

TABLE. VII. SUMMARY OF REGRESSION MODEL AND ANOVA

Source	SS	df	MS	
Model	13.933	4	3.483	Number of obs =210, F (4, 205) = 16.48 Prob > F= 0 R-squared = 0.243 Adj R-squared= 0.229, Root MSE= 0.459
Residual	43.329	205	0.211	
Total	57.263	209	0.274	

TABLE. VIII. REGRESSION COFFICIENT

Academic Impacts	Coefficient	Standard Error	T Statistics	P>t	95% Conf.Interval	
ICT in Class Room	0.025083	0.0370895	0.68	0.5	-0.048	0.0982
Outside the Class Room	-0.03696	0.0647209	-0.57	0.569	-0.1645	0.0906
Opinion Towards ICT Use	0.086031	0.0660603	1.3	0.194	-0.0442	0.2162
Addiction of ICT	0.536722	0.0766979	7	0	0.3855	0.6879
Constant	1.168289	0.1381918	8.45	0	0.8958	1.4407

E. Data Mining

In this research, Association rule mining was used, which is one of the famous data mining methods for detecting and extracting useful information from transaction data. From among several association rule mining algorithms, this research mine association rules using both the A prior and eclat algorithm in R using a package called a rule. Moreover, these methods generate a large number of available rules and make it difficult to relocate interesting ones. Therefore, this research found the one by applying different support and confidence level. At initial levels, after applying the A prior algorithm with low support and confidence value, total 1592 rules was generated. Later the test was done on different support and confidence values and stops at support value 0.07 and confidence value 0.5 and got 383 association rules. After carefully filtering the rules with confidence value 0.8, 131 quality rules were generated. Fig. 3, 4, 5 and 6 shows the Scatter and a matrix plot of the rules before and after filtering respectively. In Table X few of them with higher influence is shown. By applying eclat algorithms, total 59 rules were generated. In total 230 rows and 18 columns (details shown in Table IX and Table X), Internet_Browse_ Regularly=Yes was found to be a most frequent item and influential factor for CGPA improvement of both the algorithms.

TABLE. IX. MOST FREQUENT ITEMS

Items	Frequency
Internet_Browse_Regularly=Yes	203
Gender=Male	132
Technology=Laptop	106
cgpa1=Excellent	105
cgpa1=good	105
(Other)	607

TABLE. X. FREQUENT RULES

Sorted No	Lhs	Rhs	Support	Confidence	Lift	Count
15	{Technology=Internet, Browsing_Period=Two Hour}	{cgpa1=Excellent}	0.038	0.615	1.231	8
16	{Technology=Internet,Internet_Browse_Regularly=Yes, Browsing_Period=Two Hour}	{cgpa1=Excellent}	0.038	0.615	1.231	8
18	{Technology=Desktop,Browsing_Period=Three Hour}	{cgpa1=Excellent}	0.029	0.6	1.2	6
13	{Technology=Desktop, Internet_Browse_Regularly=Yes, Browsing_Period=Three Hour}	{cgpa1=Excellent}	0.029	0.667	1.333	6
17	{Technology=Desktop,Browsing_Period=Two Hour}	{cgpa1=Excellent}	0.014	0.6	1.2	3
19	{Technology=Desktop,Internet_Browse_Regularly=Yes,Browsing_Period=Two Hour}	{cgpa1=Excellent}	0.014	0.6	1.2	3
6	{Technology=Internet,Browsing_Period=One Hour}	{cgpa1=Excellent}	0.014	0.75	1.5	3
7	{Technology=Internet,Internet_Browse_Regularly=Yes,Browsing_Period=One Hour}	{cgpa1=Excellent}	0.014	0.75	1.5	3
9	{Internet_Browse_Regularly=No,Browsing_Period=Three Hour}	{cgpa1=Excellent}	0.01	0.667	1.333	2
10	{Technology=Laptop,Internet_Browse_Regularly=No}	{cgpa1=Excellent}	0.01	0.667	1.333	2
2	{Technology=Internet,Internet_Browse_Regularly=No}	{cgpa1=Excellent}	0.01	1	2	2
4	{Technology=Internet,Internet_Browse_Regularly=No,Browsing_Period=Three Hour}	{cgpa1=Excellent}	0.01	1	2	2
11	{Technology=Laptop,Browsing_Period=Two Hour}	{cgpa1=good}	0.076	0.667	1.333	16
8	{Technology=Laptop,Internet_Browse_Regularly=Yes,Browsing_Period=Two Hour}	{cgpa1=good}	0.071	0.682	1.364	15
14	{Internet_Browse_Regularly=Yes,Browsing_Period=One Hour}	{cgpa1=good}	0.038	0.615	1.231	8
12	{Technology=Laptop,Internet_Browse_Regularly=Yes,Browsing_Period=One Hour}	{cgpa1=good}	0.019	0.667	1.333	4
3	{Technology=Desktop,Browsing_Period=One Hour}	{cgpa1=good}	0.014	1	2	3
5	{Technology=Desktop,Internet_Browse_Regularly=Yes,Browsing_Period=One Hour}	{cgpa1=good}	0.014	1	2	3
1	{Browsing_Period=Zero Hour}	{cgpa1=good}	0.01	1	2	2

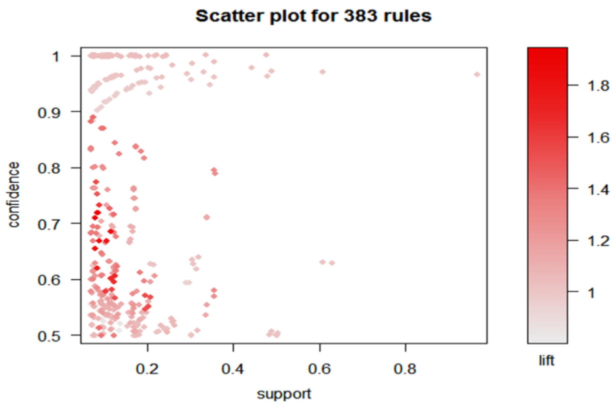


Fig. 3. Scatter Plot of Rules before Filtering.

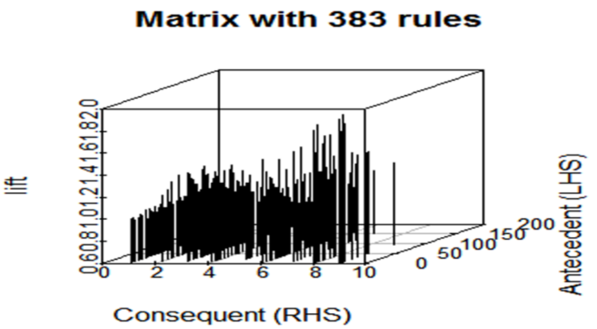


Fig. 4. Matrix Plot of Rules before Filtering.

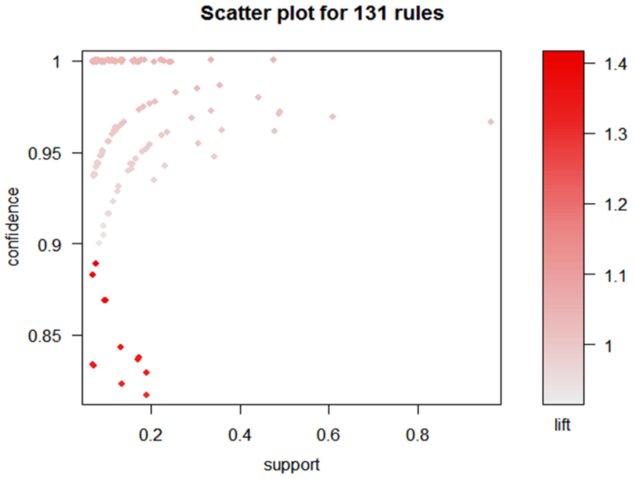


Fig. 5. Scatter Plot of Rules after Filtering.

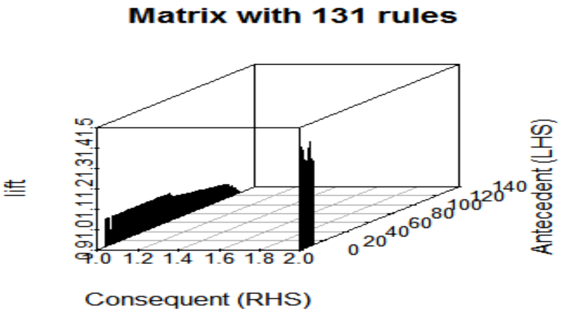


Fig. 6. Matrix Plot of Rules after Filtering.

V. Comparison

This research compares the OLS regression analysis and path structural equation modeling. Although path analysis is an extension of linear regression model used to examine the relationships between measured variables, but it is a highly flexible and extensible methodology. Tests associated with both methods are satisfied with their assumption. The OLS regression and path structural equation model identify different relationships between variables in the model. Both models include the same predictor and predicted variables in different items. Although statistical tests of significance and an R-square value slightly differ, but the variable “Addiction of ICT” is the main significant comparative predictor in both models. The result from association rule mining also reflects the same, that is, if the student increases the use of ICT then their CGPA improves provided that, they are using ICT in academic purpose.

VI. Discussion

The findings of the study show that the student ‘Addiction of ICT’ is the most important predictor in the analysis. This gives the controversial result with the academic impact of ICT. This result also supports the low value of R2 (0.345 and 0.243) in the studied models. From these findings, drawn conclusions are that the student academic result is not only depending on ICT but also has a lot of external and internal factors. Moreover, Students spent most of the time with ICT in non-academic purpose.

VII. Recommendations

The authors recommend the following to improve the students’ academic performance based on the findings of this study:

- ICT facilities in classrooms should be improved
- Teachers should conduct their classes by using ICT
- Technology should be used for own advancement and should control the unnecessary use of technology.
- Students should use ICT for their academic purposes most of the time.
- All the university should adopt the technology for academic purposes.
- Students should make aware of ICT use in Education

VIII. Conclusion and Future Works

In this era, ICT plays an important role in day- to-day activities, including education, so it is high time to evaluate the impact of ICT on education and to ensure its positive use. In this context, this study was carried out using Descriptive statistics, reliability analysis, Confirmatory Factor Analysis, OLS regression, path analysis, and data mining algorithms such as Association rule mining and eclat. The study exposes the negative impact of the academic performance of the students as the use of ICT if it is not used properly. The study also shows that if academic or related institutions take the right steps to use the ICT for academic purposes, education as a whole and students ’ academic performance in particular will benefit greatly. This study will be extended in future to compare the impact of ICT on the performance of students in private universities with that of students in the public university of the country. The data set will also be enlarged and the prediction from this data set will be included.

References

[1] J. AlAmmary, "Educational Technology: A Way to Enhance Student Achievement at the University of Bahrain", Procedia-Social and Behavioral Sciences, vol. 55, pp. 248-257, 2012.

[2] M. Alderete and M. Formichella, "The effect of ICTs on academic achievement: The conectar igualdad programme in Argentina", CEPAL Review, vol. 2016, no. 119, pp. 83-100, 2017.

[3] A. Aristovnik, "The Impact of ICT on Educational Performance and its Efficiency In Selected EU and OECD Countries: A Non-Parametric Analysis", SSRN Electronic Journal, 2012.

[4] Aziz, A. A., Idris, W. M. R. W., Hassan, H., Jusoh, J. A., and Emran, N. A., “Implementing Aproiri Algorithm for Predicting Result Analysis”, GSTF Journal on Computing (JoC), vol.2, no.4, 2018.

[5] W. Basri, J. Alandejani and F. Almadani, "ICT Adoption Impact on Students’ Academic Performance: Evidence from Saudi Universities", Education Research International, vol. 2018, pp. 1-9, 2018.

[6] Y. Baek, J. Jung and B. Kim, "What makes teachers use technology in the classroom? Exploring the factors affecting facilitation of technology with a Korean sample", Computers & Education, vol. 50, no. 1, pp. 224-234, 2008.

[7] Beggs, T.A. (2000). Influences and barriers to the adoption of instructional technology, retrieved April 25, 2008, from ht tp://www.mtsu.edu/~itconf/proceed00/beggs/beggs.htm.

[8] J. van Braak, "Factors influencing the use of computer mediated communication by teachers in secondary schools", Computers & Education, vol. 36, no. 1, pp. 41-57, 2001.

[9] Butler, D. L. and Sellbom, M., “Barriers to adopting technology”, Educause Quarterly, vol. 2, pp. 22-28, 2002.

[10] J. Bussey, T. Dormody and D. VanLeeuwen, "Some Factors Predicting the Adoption of Technology Education in New Mexico Public Schools", Journal of Technology Education, vol. 12, no. 1, 2000.

[11] A. Carle, D. Jaffee and D. Miller, "Engaging college science students and changing academic achievement with technology: A quasi-experimental preliminary investigation", Computers & Education, vol. 52, no. 2, pp. 376-380, 2009.

[12] G. Conole, M. de Laat, T. Dillon and J. Darby, "‘Disruptive technologies’, ‘pedagogical innovation’: What’s new? Findings from an in-depth study of students’ use and perception of technology", Computers & Education, vol. 50, no. 2, pp. 511-524, 2008.

[13] A. Enriquez, "Enhancing Student Performance Using Tablet Computers", College Teaching, vol. 58, no. 3, pp. 77-84, 2010.

[14] C. Fried, "In-class laptop use and its effects on student learning", Computers & Education, vol. 50, no. 3, pp. 906-914, 2008.

[15] J. Hair, B. Babin and N. Krey, "Covariance-Based Structural Equation Modeling in the Journal of Advertising: Review and Recommendations", Journal of Advertising, vol. 46, no. 3, pp. 454-454, 2017.

[16] M. Islam and M. Fouji, “The impact of ICT on students’ performance: A case study of ASA University Bangladesh”, ASA University Review, vol. 4, no. 2, pp. 101-106, 2010.

[17] K.Mumcu and K. Usluel, “Mesleki ve teknik okul ogretmenlerinin bilgisayar kullanımları ve en- geller”, Hacettepe Universitesi Egitim Fakultesi Dergisi, vol. 26, pp.91- 100, 2004.

[18] Nketiah-Amponsah, E., Asamoah, M. K., Allassani, W., &Aziale, L. K. (2017). Examining students’ experience with the use of some selected ICT devices and applications for learning and their effect on academic performance. Journal of Computers in Education, 4(4), 441-460.

[19] Perbawaningsih, Y. (2013). Plus minus of ICT usage in higher education students. Procedia-Social and Behavioral Sciences, 103, 717-724.

[20] Roscoe, J. T. (1975). Fundamental research statistics for the behavioral sciences [by] John T. Roscoe.

[21] Sekaran, U. and Bougie, R., (2010). Theoretical frame- work In theoretical framework and hypothesis development. Research methods for business: A skill building approach, 80.

[22] Sari, A. (2014). Influence of ICT Applications on Learning Process in Higher Education. Procedia—Social and Behavioral Sciences, 116, 4939-4945.

[23] Sife, A., Lwoga, E., &Sanga, C. (2007). New technologies for teaching and learning: Challenges for higher learning institutions in developing countries. International journal of education and development using ICT, 3(2), 57-67.

[24] S.Smith, and J. Caruso, “The ECAR study of undergraduate students and information technology”, 2010.

[25] Song, H. D., & Kang, T. (2012). Evaluating the Impacts of ICT Use: A Multi-Level Analysis with Hierarchical Linear Modeling. Turkish Online Journal of Educational Technology- TOJET, 11(4), 132-140.

[26] Turel, Y. K. (2011). An interactive whiteboard student survey: Development, validity and reliability. Computers & Education, 57(4), 2441-2450.

[27] Ward, L., & Parr, J. M. (2010). Revisiting and reframing use: Implications for the integration of ICT. Computers and Education, 54(1), 113-112.

[28] Yamamoto, K. (2007). Banning laptops in the classroom: Is it worth the hassles? Journal of Legal Education, 57(4), 477-520.

[29] Youssef, A. B., &Dahmani, M. (2008). The impact of ICT on student performance in higher education: Direct effects, indirect effects and organisational change. RUSC. Universities and Knowledge Society Journal, 5(1), 45-56.

[30] Youssef, A. B., & DAHMANI, M. Innovative ICT usage by Higher Education Teachers in Tunisia: A PLS Path Modelling Approach.

[31] Youssef, A. B., Youssef, H. B., &Dahmani, M. (2013). Higher education teachers e-skills and the innovation process. International Journal of Computer and Information Technology, 2(2),185-19.

[32] Joseph F. Hair Jr., Barry J. Babin & Nina Krey (2017) Covariance-Based Structural Equation Modeling in the Journal of Advertising: Review and Recommendations, Journal of Advertising, 46:1, 163-177, DOI: 10.1080/00913367.2017.1281777.

WhatsApp as an Educational Support Tool in a Saudi University

Ahmad J Reeves[1], Salem Alkhalaf[2]
Computer Department, Qassim University
Alrass, Saudi Arabia

Mohamed A. Amasha[3]
Department of Computer Teacher Preparation
Damietta University, Egypt

***Abstract*—WhatsApp is a widely used social media app, growing in popularity across the Middle East, and the most popular in Saudi Arabia. In this paper, we investigate the usage of WhatsApp as an educational support tool in a Saudi university. An online survey was constructed to ascertain how students and staff feel about and utilize WhatsApp as part of their daily studies. It also aimed to gather their thoughts on other platforms offered by the university such as Blackboard and email. The survey was tested and the results analyzed for frequency distributions, mean score, and standard deviation. Our results from nearly 200 student and staff members reveals that WhatsApp is heavily utilized for a variety of educational support tasks and greatly preferred over the other platforms. We propose that WhatsApp has good potential to support not only student coordination, information dissemination and simple enquiries but also to support formal teaching and out-of-class learning.**

Keywords—Online learning; WhatsApp; e-learning; blackboard; communication; mobile learning

I. Introduction

WhatsApp is one of the most widely used social messaging platforms in the world, with over one billion people now using the service [1]. WhatsApp allows users to freely share text and audio messages (along with a range of other media), and form groups related to mutually interesting topics. In other words, WhatsApp offers a useful, lightweight communication tool on a mobile platform. WhatsApp and other messaging apps have become embedded into daily life for millions of people, with nearly half of smartphone owners in the US using WhatsApp or KIK [2]. The total number of WhatsApp users has also increased to 1.2 billion (see Fig. 1).

Compared to Facebook, WhatsApp messages remain private (as there are fewer privacy policy changes than Facebook), security is end-to-end on WhatsApp and response times on WhatsApp are quicker with instant notification of receipt and whether the message has been read. WhatsApp is also the top social media app in Saudi Arabia used by 56% of the population [4]. Given this popularity, in this paper, we focus on how WhatsApp is also being used as an educational support tool in a Saudi University. We want to compare how students are using WhatsApp in comparison to an existing mobile Learning Management System (LMS) and e-mail. Anecdotal evidence suggested that students used the LMS and e-mail very rarely and were more inclined to rely on social media platforms as support tools during their studies. We wished to find the extent of this usage and any potential for other educational uses.

In section two, we present a literature review of how WhatsApp has been used in educational settings and the effects (both positive and negative) on that education. In section three, we present the methodology for the study. In section four, we present the results and in sections five and six the discussion and conclusions.

II. Literature Review

Studies have shown that the use of WhatsApp in educational settings have produced either a positive effect, a negative effect or a mixture of both, either on the standard of education or the students themselves.

A. Positive Effects

Amry [5] compared teaching using WhatsApp for a unit at university with face-to-face instruction for 15 students. Results indicated the WhatsApp group had a more positive achievement for the unit undertaken and more positive attitudes to WhatsApp, specifically noting its ability to make learning easy, improved problem solving and knowledge sharing. Patil, Depthi & Tadasad [6] in their study of WhatsApp use amongst 94 postgraduate university students found that approximately half of them used the platform to share academic information alongside their regular communications.

WhatsApp has also been found to have a positive effect when used in a blended mobile lecture environment. Barhoumi [7] reported several benefits of student usage of WhatsApp including increased discussion, collaboration and document sharing. Minhas, Ahmed & Ullah [8] also found a positive effect of WhatsApp. In their survey of 84 university students, approximately one fifth used it for educational purposes alongside regular friends and family communication.

The previous studies focused on the student's view of WhatsApp. Gachago, Strydom, Hanekom, Simons & Applters [9] studied three lecturer's perceptions of using WhatsApp for distance learners and an on-campus course. They found that WhatsApp helped in the facilitation and coordination of learning, with the mobility aspect 'blurring the physical and geographical boundaries'. The social nature of messaging on WhatsApp also, they noted, led to learners and facilitators engaging in more informal ways 'crossing professional and social boundaries'. Although this blurring had positive outcomes, some negative aspects reported included increased stress and a lack of privacy, solved by negotiated ground rules initiated by both sides.

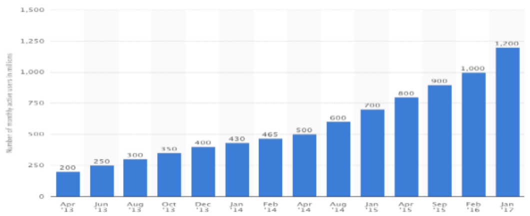


Fig. 1. Monthly Active WhatsApp users Worldwide from 2013 to 2017 [3].

B. *Negative Effects*

Yeboah & Ewur [10] investigated the effect of WhatsApp on the performance of students in schools in Ghana. Fifty students were interviewed and five hundred questionnaires were taken. Apart from ease of communications and productive sharing of information, they found many undesirable effects of using WhatsApp including disruption from studies and incomplete assignments, poorer spelling and grammar and poor focus in lectures. Alosaimi, Alyahya, Alshahwan, Mahyijari & Shaik [11] found other negative effects relating to the health of students using smartphones. They found that more than a quarter of the 2367 students they surveyed used their mobiles for more than 8 hours per day, leading to less sleeping time, decreased energy, an unhealthier lifestyle and over a quarter reported a negative effect on their academic achievement.

C. *Both Positive and Negative Effects*

Ahad & Lim [12] in their study of WhatsApp use of 158 undergraduates found the benefits of WhatsApp included sharing academic or study related information alongside regular communications with family and friends. Negative aspects they reported included addictive like behavior in checking messages (having a negative impact on their studies) and the spread of false or unregulated information.

Bouhnik and Deshen [13] interviewed twelve instructors who had utilised WhatsApp to connect with groups they taught in high school. They reported that the WhatsApp groups are used for:

- communicating with students;
- encouraging the social atmosphere;
- building dialogue and boosting sharing among students;
- as a learning platform.

Their students highlighted both technical advantages of WhatsApp (simplicity, low cost and accessibility) and educational advantages (pleasant learning environment, the availability of learning materials, teacher accessibility, and the furtherance of learning beyond class hours). On the negative side, they also reported that instructors found the high volume of irrelevant messages annoying, differences of languages between students and privacy problematic as it was assumed teachers should be available 24 hours a day.

D. *WhatsApp in Saudi Arabia*

As this study takes place in Saudi Arabia (KSA), it is important to contextualize how WhatsApp is used in KSA in order to compare with the previous literature. Abdulkareem [14] in a study of Saudi middle school science teachers and students found that all teachers had mobile phones and all used WhatsApp as the main social media tool. 75% of their students owned a mobile phone and 73% of students also used WhatsApp. He also found a willingness amongst both teachers and students to use WhatsApp in education, only restricted by educational infrastructure and training using this platform. In a survey of over 500 students at King Abdul-Aziz University, Aifan [15] found a positive attitude towards using social media in education with the most frequent tool used by students being WhatsApp. The only concerns raised were privacy and regulation of inappropriate content. Alqahtani [16] and Alsurehi & Youbi [17] also found that social networking technologies might be implemented into Saudi education without major obstacles. WhatsApp was also found to have a positive affect when used to support writing skill improvement.

However, other studies have shown that even given this large usage of WhatsApp, the quality of the communication taking place can be poor. Al Lily [18] in a study of communicative content of social media messages coined the term ‘information thinness’ to mean the lack of information or trivial information in messages. His survey of the content of messages sent by over 700 people in KSA found that:

> ‘Users may overuse technology merely to communicate for the sake of communication, with no interest in actually being informed by anything, thus suffering from ‘information thinness’.

The previous studies have highlighted that the use of WhatsApp in educational settings have produced a variety of both positive and negative effects. In our study, we wanted to find out how WhatsApp has been used both educationally and in comparison to an existing LMS and email.

III. Methodology

An online survey was produced and the link distributed to both female and male students and staff in the Department of Computer Science at Qassim University in Saudi Arabia. The survey was designed to look at students’ attitudes toward the use of WhatsApp in education. All participants were requested to complete the online survey via the website (https://goo.gl/forms/6o73kjHNOlsLKms43). They were provided with information about the survey and volunteered to participate after reading the importance of the study. The survey utilized a 5-point Likert scale: (1= strongly disagree, 2= disagree, 3= slightly agree, 4= agree, 5= strongly agree). It comprised 19 closed-ended questions organized in four sections: General Usage of WhatsApp (GU); Education Use of WhatsApp (EU); Other Educational Tools (OE); and Potential of WhatsApp in Education (PE). Students and staff were requested to click the Google URL to complete the survey (N=196). Google documents was used to collect the resultant data in spreadsheet form. The questionnaire was validated by 10 experts, and its internal reliability was found to be good. The Cronbach’s α coefficient was found to be 0.745.

A. *Data Analysis*

The resultant data was analyzed using SPSS statistical software package V.20.0. The frequency distributions, mean score, and standard deviation were computed for each item from the questionnaires in the Google spreadsheet. A Chi-square test was utilised to equate between actual and potential students' responses regarding the use of WhatsApp in education. The statistical significance level was fixed at p<0.05.

B. *Participants*

The study at Qassim University in KSA was conducted during the first semester of the 2016-2017 academic year. Participants were selected using a systematic random method. As shown in Table I below, the number of male students was 58 (29.6%) and the number of female students was 138 (70.4%). Participants' ages ranged from 22-25 (M=21.92). In addition, the demographic questionnaire indicated that students numbered 178 (90.8%) and the number of lecturers was 18 (9.2%) (see Table I).

TABLE. I. DEMOGRAPHIC INFORMATION

Items	Frequency (n=196)	(%)
Gender		
Male	58	29.6
Female	138	70.4
Age		
18-21	65	33.2
22-25	80	40.8
26-29	28	14.3
30-33	7	3.6
34+	16	8.2
Are you a student or lecturer?		
Student	178	90.8
Lecturer	18	9.2

IV. RESULTS

A. *General Usage of WhatsApp (GU)*

The results in (Table II and Fig. 2) indicate that almost all participants (N = 160; 81.6 %) reported that they use WhatsApp instant messaging on their phones. There was a significant difference between participated (male, female) regarding using WhatsApp, (χ2 (1, N = 196) = 78.44, p < .01). This highlights that the difference between the observed frequencies and the expected ones in relation to the question raised is statistically significant and not due to chance factor.

In (Table III and Fig. 3) above, summarizes the results of student's opinion about the length of time using WhatsApp identified in the analysis. These themes all relate to students' views on how long they use WhatsApp. Almost all students (82.7%) use WhatsApp for two years (M =4.74, X2=488.13).

TABLE. II. USAGE OF WHATSAPP INSTANCE MESSAGING

N	Question	Yes		No		*M*	*Sd*	X^2
		f	%	F	%			
1	Do you use WhatsApp instant messaging on your phone? (Y)/(N)	160	81.6	36	18.4	1.18	0.449	78.44

* p>0.01. ** p>0.001.

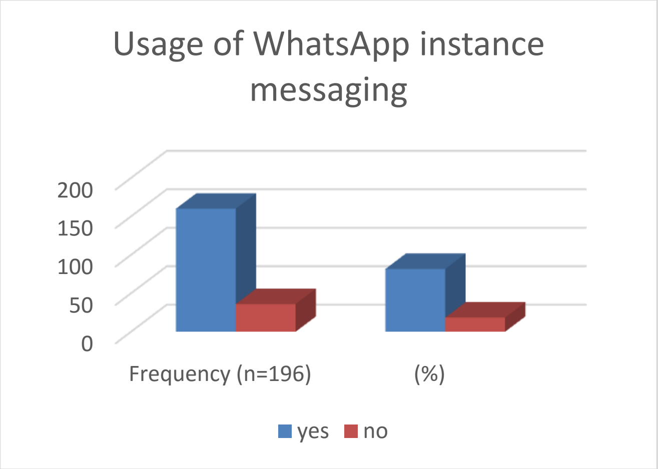


Fig. 2. Usage of WhatsApp Instance Messaging.

TABLE. III. LENGTH OF TIME USING WHATSAPP

Items	Frequency (n=196)	(%)	*M*	*Sd*	X^2
How long have you been using WhatsApp?			4.75	0.611	488.13^{**}
Less than 1 year	-	-			
1 year	22	11.2			
2 year	162	82.7			
3 year	9	4.6			
More than 4 years	3	1.5			

* p>0.01. ** p>0.001

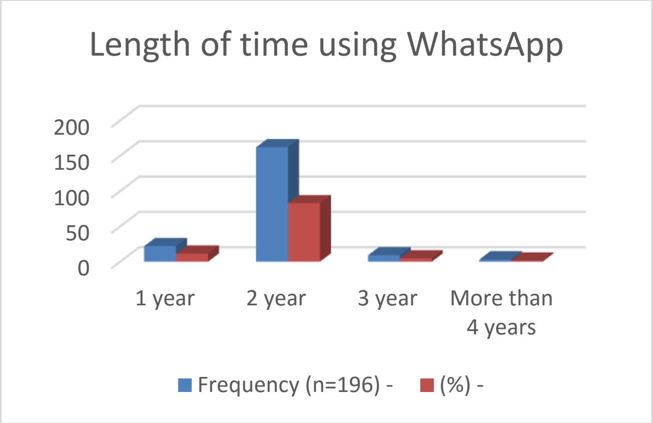


Fig. 3. Length of Time using WhatsApp.

The above question (Table IV and Fig. 4) shows that approximately half of the participants 58.7% reported that they use WhatsApp in their academic time with both lecturers and with other members of college staff (M =2.99, X2=218.28). On the other hand, 21.9% used for contact with their friends. This result indicates that students are interested in using WhatsApp for an educational purpose.

In terms of content sent using WhatsApp and what students use WhatsApp with (Table V), the majority reported that they use it in all type of the content (81.1%). The participants reported that they almost never use WhatsApp just in text messages (15.3%). All responses are shown in Fig. 5.

TABLE. IV. CONTACT USING WHATSAPP

Who do you contact using WhatsApp?	Frequency (n=196)	(%)	*M*	*Sd*	X^2
Friends	43	21.9	2.99	1.307	218.28^{**}
Family members	34	17.3			
Other students	1	0.5			
Lecturers & members of college staff	115	58.7			
Work colleagues	3	1.5			

* p>0.01. ** p>0.001

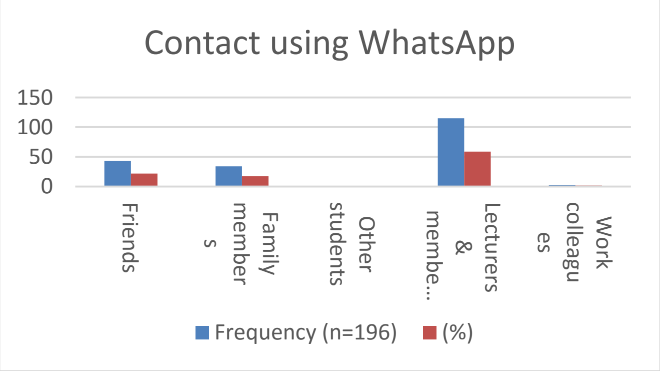


Fig. 4. Contact using WhatsApp.

TABLE. V. CONTENT SENT USING WHATSAPP

What type of content do you send?	Frequency (n=196)	(%)	*M*	*Sd*	X^2
Text messages	3	15.3	4.33	1.45	472.41^{**}
Videos	1	0.5			
Audio recordings	2	1.00			
Images	4	2.00			
Documents (e.g. .pdf/.doc/web link/maps)	0	0			
All	159	81.1			

* p>0.01. ** p>0.001

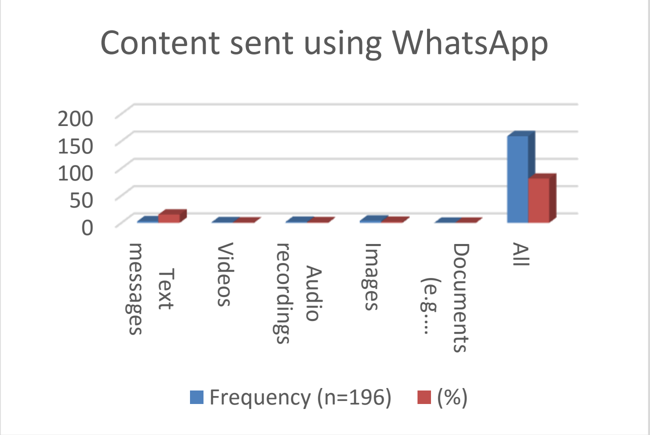


Fig. 5. Content Sent using WhatsApp.

Most students used WhatsApp for about 3-5 hours per day (see Table VI and Fig. 6). The results indicate that (M =2.74, X2=30.78**) X2 has high statistical significance.

B. *Education Use of WhatsApp (EU)*

In Table VII, students reported that they used WhatsApp for text messages. We asked them about the last 7 days, how many of their WhatsApp messages do they think relate to their studies at University. Approximately half of the participants (45.1%) reported more than 60 of their messages were related to their studies.

TABLE. VI. HOURS PER DAY USING WHATSAPP

How many hours per day do you spend using WhatsApp?	Frequency (n=196)	(%)	*M*	*Sd*	X^2
Less than 1 hour	34	17.3	2.74	1.24	30.78^{**}
1-2 hours	55	28.1			
3-5 hours	60	30.6			
6-8 hours	22	11.2			
More than 8 hours	25	12.8			

* p>0.01. ** p>0.001

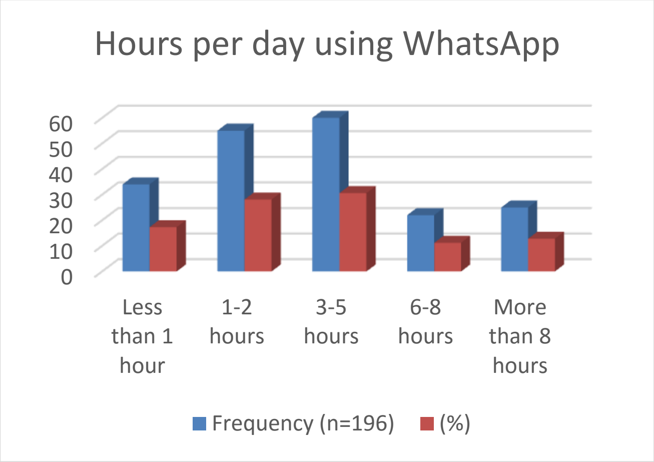


Fig. 6. Hours Per Day using WhatsApp.

TABLE. VII. NUMBER OF MESSAGES

Items	Frequency (n=196)	(%)	*M*	*Sd*	X^2
In the last 7 days how many of your WhatsApp messages do you think relate to your studies at University?					
None	14	7.6	79.69	3.68	1.410^{**}
1-20	37	20.1			
21-40	25	13.6			
41-60	25	13.6			
More than 60	83	45.1			

* p>0.01. ** p>0.001

In Table VIII and Fig. 7 can be seen that the majority of students (74 out of 160, 37.8%) used WhatsApp in groups (3 or more). This result confirms the attractiveness of students to use social networks.

TABLE. VIII. WHATSAPP GROUPS

How many WhatsApp groups are you a part of that relate to your studies?	Frequency (n=196)	(%)	*M*	*Sd*	*X²*
None	15	7.7	3.7	1.22	66.19**
1	22	11.2			
2	27	13.8			
3	74	37.8			
4 or more	58	29.6			

*p>0.01. ** p>0.001.

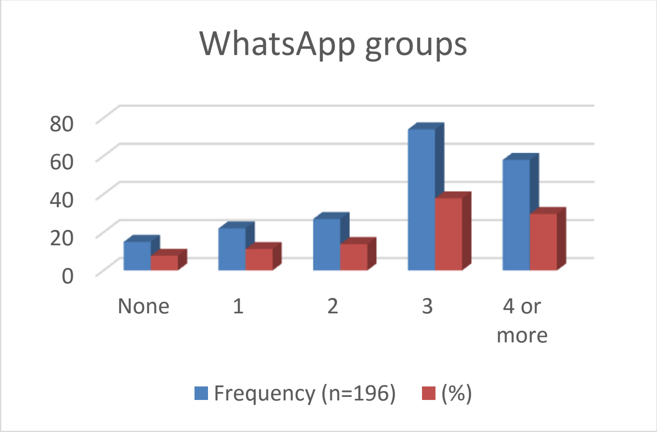


Fig. 7. WhatsApp Groups.

* p>0.01. ** p>0.001.

Questionnaire data showed that all students used WhatsApp to exchange information regarding lecture times & locations, questions with students, subject content, deadlines and exams and questions with a lecturer. The results in (Table IX and Fig. 8), showed high statistical significance (M =5.37, X2=87.12**).

According the results in (Table X and Fig. 9), the students feel uncomfortable with the whole range of privacy issues whilst using WhatsApp.

TABLE. IX. TYPES OF INFORMATION SENT

If you use WhatsApp groups for your studies, what type of information is exchanged?	Frequency (n=196)	(%)	*M*	*Sd*	*X²*
Lecture times & locations	25	12.8	5.37	1.40	87.12**
Questions with students	6	3.1			
Subject content	4	2.0			
Deadlines and exams	1	.5			
Questions with lecturer	0	0			
All	160	81.6			

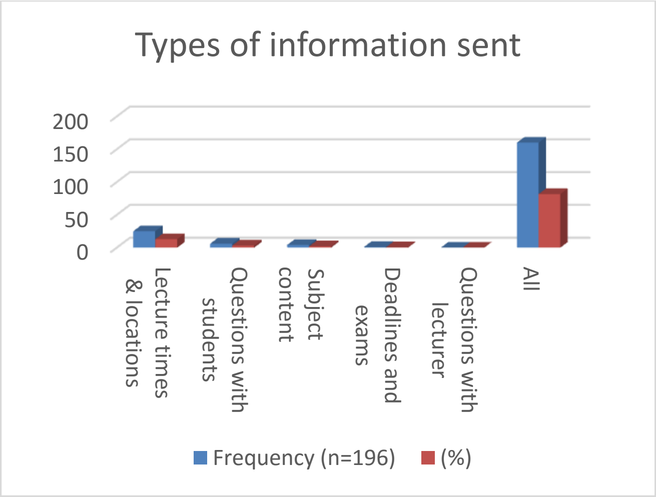


Fig. 8. Types of Information Sent.

TABLE. X. PRIVACY ISSUES

Which privacy issues do you feel uncomfortable with using WhatsApp?	Frequency (n=196)	(%)	*M*	*Sd*	*X²*
Giving out your phone number	51	26.0	3.72	1.983	38.662**
Using a profile picture	8	4.1			
Discussing non-educational related topics	31	15.8			
Sharing personal information	17	8.7			
Last seen or status information	32	16.3			
All	57	29.1			

* p>0.01. ** p>0.001.

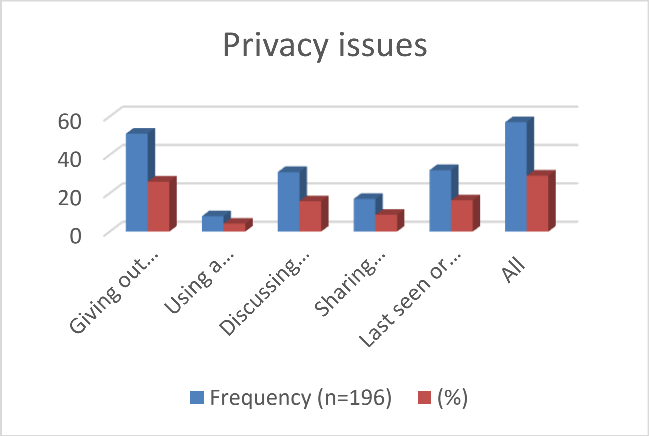


Fig. 9. Privacy Issues.

C. *Other Educational Tools (OE)*

Furthermore, the majority of our respondents (68 out of 160) never use the existing Blackboard system (see Table XI and Fig. 10). All these results are important when considering the implementation of WhatsApp in their work. Students prefer to use WhatsApp for their learning management system. This indicates that social networks are becoming a more common and friendly platform.

TABLE. XI. USAGE OF BLACKBOARD

How often do you use the existing online Blackboard System?	Frequency (n=196)	(%)	*M*	*Sd*	X^2
Never	68	34.7	2.13	1.047	90.27^{**}
Once or twice per month	56	28.6			
Once or twice per week	58	29.6			
Every day	7	3.6			
More than once per day	7	3.6			

* p>0.01. ** p>0.001.

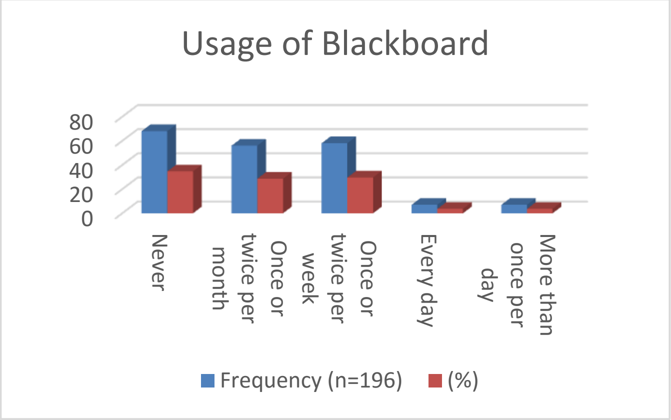


Fig. 10. Usage of Blackboard.

Furthermore, the majority of our samples (66 out of 160, 33.7%) never use e-mail. This result is important when considering the implementation of social media (WhatsApp) in their work. Students prefer to use social media (WhatsApp) for their learning management system. This indicates that social networks are becoming more a more common and preferred platform (see Table XII).

D. Potential of WhatsApp in Education (PE)

According to the results in (Table XIII), the majority of students (n=176, 90.25%) reported that WhatsApp is very easy to use.

As indicated by the parameter estimates and results in (Table XIV and Fig. 11), our respondents reported (n=144, 75.5%) that WhatsApp was very useful in supporting their studies and learning.

TABLE. XII. USAGE OF EMAIL

How often do you use e-mail in relation to your studies?	Frequency (n=196)	(%)	*M*	*Sd*	X^2
Never	66	33.7	2.24	1.082	66.45^{**}
Once or twice per month	46	23.5			
Once or twice per week	54	27.6			
Every day	30	15.3			
More than once per day	0	0			

* p>0.01. ** p>0.001.

TABLE. XIII. EASE OF USE

Overall, how easy do you find WhatsApp to use?	Frequency (n=196)	(%)	*M*	*Sd*	X^2
Very easy to use	176	90.25	4.79	.641	607.58^{**}
Easy to use	-	-			
Neither easy or difficult to use	18	9.23			
Difficult to use	-	-			
Very difficult to use	1	.512			

* p>0.01. ** p>0.001.

TABLE. XIV. USEFULNESS OF WHATSAPP

How useful do you find WhatsApp in supporting your studies?	Frequency (n=196)	(%)	*M*	*Sd*	X^2
Not useful	3	1.5	4.48	.947	414.66^{**}
Some use	-	-			
Good use	45	23.0			
Very useful	148	75.5			
Extremely useful	--	-			

* p>0.01. ** p>0.001.

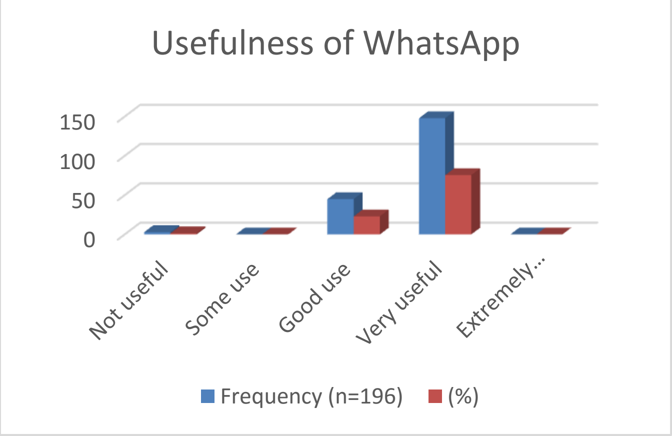


Fig. 11. Usefulness of WhatsApp.

The results shown in Table XV and Fig. 12 indicated that WhatsApp has excellent potential for teaching and learning. 37.8% of participants reported that the WhatsApp has good potential and 31.1% of participants reported that WhatsApp had excellent potential.

TABLE. XV. POTENTIAL OF WHATSAPP

Apart from supporting your studies, how much potential do you think WhatsApp has for actual teaching and learning?	Frequency (n=196)	(%)	*M*	*Sd*	X^2
None	13	6.6	3.82	1.125	83.796^{**}
Minimal potential	9	4.6			
Some potential	39	19.9			
Good potential	74	37.8			
Excellent potential	61	31.1			

* p>0.01. ** p>0.001.

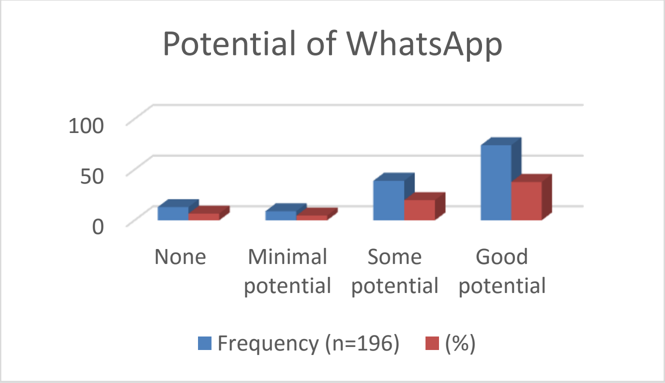


Fig. 12. Potential of WhatsApp.

WhatsApp was reported to have a positive effect on the education process by 95.4% of the students (Table XVI).

The majority of students reported that they preferred using WhatsApp to Blackboard. See Table XVII and Fig. 13.

TABLE. XVI. POSITIVE/NEGATIVE EFFECT OF WHATSAPP ON EDUCATION

Overall, do you feel WhatsApp has a positive or negative affect on your education?	Frequency (n=196)	(%)	*M*	*Sd*	X^2
Positive	187	95.4	1.05	.210	698.13**
Negative	9	4.6			

* p>0.01. ** p>0.001.

TABLE. XVII. OVERALL PREFERENCE OF WHATSAPP OR BLACKBOARD

If you had to choose between using either WhatsApp or Blackboard to support your studies, which would you choose?	Frequency (n=196)	(%)	*M*	*Sd*	X^2
Blackboard	29	14.8	1.85	.356	536.90**
WhatsApp	167	85.2			

* p>0.01. ** p>0.001.

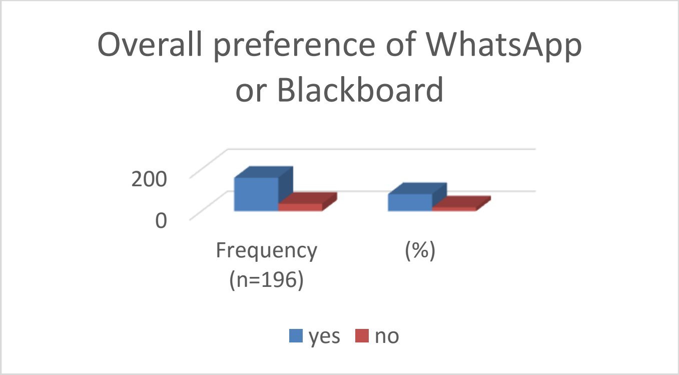


Fig. 13. Overall Preference of WhatsApp or Blackboard.

V. DISCUSSION

The results of our study have highlighted how widely WhatsApp has become the de-facto platform chosen by these students for both educational organization and information dissemination. With the high level of mobile phone usage in Saudi Arabia, it is evident that the educational sphere needs to understand and develop ways to exploit this phenomenon in order to support the potential that such social media channels provide. Our study has supported the previous findings that WhatsApp increases and supports both document sharing and coordination of educational information (see [9], [8] & [15]). Our study also highlighted not only high levels of student-to-student contact, but also student to lecturers and other members of staff. With nearly half of all messages education related, these channels (both in and out of groups) supported both students requirements for timely information but also lecturer dissemination of that information.

In comparison to the other existing platforms the university offers, WhatsApp was greatly preferred over Blackboard and e-mail (which were hardly made use of). This may be due to the convenience of using a lightweight, always on mobile platform, over a desktop-based login system. Indeed, even though Blackboard and email are available on mobile devices, they are still not preferred methods. In could be the accessibility, usability and relative simplicity of WhatsApp makes it a first choice for students, only reverting to other existing platforms when necessary. It is also possible that students prefer to have both their social and educational messages in one platform rather than three. However, our study has reiterated how privacy issues remain an important issue with relation to WhatsApp, something future educational designers will have to consider carefully.

With regard to teaching and learning, our study has shown that WhatsApp has excellent potential to transform into not just an organizational tool but also a mainstream teaching tool. Future work will need to look at potential methods and designs that could support teaching and learning through WhatsApp beyond simple information dissemination and clarifications from members of staff. With 95% of respondents indicating that WhatsApp had a positive effect on education there is high potential for future research to see how this can be implemented.

Although our study was limited to one university, future work that assesses this trend across all universities in the Kingdom would provide a more detailed and nuanced view of both the level of social media usage and reasons for any preferences.

VI. CONCLUSIONS

Social media platforms have been widely adopted and are continuously growing in Saudi Arabia for both social and leisure activities. This study has supported the view that WhatsApp is not only a part of this trend, but has high potential to support some of the educational needs of university students. Future work in this area need to find ways for lecturers to complement their face-to-face lectures with convenient on-line learning via WhatsApp, and also for students to maximize their learning potential from high levels of mobile screen time.

References

[1] Coleman, E., & O'Connor E. The role of WhatsApp in medical education; a scoping review and instructional design model. In BMC Medical Education volume 19, Article number: 279 (2019).

[2] Alqahtani, M., Bhaskar,C., Elumalai, K.V., & Abumelha, M. Whatsapp: An online platform for University-level English language education. Arab World English Journal (AWEJ) Volume 9, Number 4, December 2018.

[3] Statista (2017). https://www.statista.com/statistics/260819/number-of-monthly-active-WhatsApp-users/.

[4] Alanzi,T., Bah, S., Alzahrani, S., Alshammari, S. & Almunsef, F. Evaluation of a mobile social networking application for improving diabetes Type 2 knowledge: an intervention study using WhatsApp. Journal of Comparative Effectiveness Research 2018.

[5] Amry, A.B. The impact of WhatsApp mobile social learning on the achievement and attitudes of female students compared with face to face learning in the classroom. European Scientific Journal August 2014 edition vol.10, No.22 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431.

[6] Patil, S, Depthi & Tadasad, P.G. Usage of WhatsApp messenger amongst post-graduate students in a university environment: A study of Karnataka state women's university, Vijayapura. International Journal of Multidisciplinary Research and Development Volume 2; Issue 11; November 2015; Page No. 591-594.

[7] Barhoumi, C. The Effectiveness of WhatsApp Mobile Learning Activities Guided by Activity Theory on Students' Knowledge Management. Contemporary Educational Technology, 2015, 6(3), 221-238.

[8] Minhas, S., Ahmed, M & Ullah, Q.F. Usage of WhatsApp: A Study of University Of PeshaWhatsAppr, Pakistan. International Journal of Humanities and Social Science Invention ISSN (Online): 2319 – 7722, ISSN (Print): 2319 – 7714 www.ijhssi.org ||Volume 5 Issue 7 ||July. 2016 || PP.71-73.

[9] Gachago, D., Strydom, S., Hanekom, P., Simons, S., & WhatsApplters, S. Crossing boundaries: lecturers' perspectives on the use of WhatsApp to support teaching and learning in Higher Education. Progressio, 2015, 37*(1)*, 172–187.

[10] Yeboah, J., & Ewur, G.D. The impact of WhatsApp messenger usage on students performance in Tertiary Institutions in Ghana. Journal of Education and Practice, 2014, *5*(6), 157-164.

[11] Alosaimi. F, Alyahya. H, Alshahwan. H, Al Mahyijari. N & Shaik. S.A. 'Smartphone addiction among university students in Riyadh, Saudi Arabia'. Saudi Medical Journal 2016; Vol. 37 (6): 675-683 doi: 10.15537/smj.2016.6.14430.

[12] Ahad, D.A Lim, S.M.A. Convenience or Nuisance?: The 'WhatsApp' Dilemma. Procedia - Social and Behavioral Sciences 155; 2014, pp.189 – 196.

[13] Bouhnik, D., & Deshen, M. WhatsApp goes to school: Mobile instant messaging between teachers and students. Journal of Information Technology Education: Research, 2014, 13, 217-231.

[14] Abdulkareem, S.A. Exploring the Use and the Impacts of Social Media on Teaching and Learning Science in Saudi. Procedia - Social and Behavioral Sciences 182 (2015) 213 – 224.

[15] Aifan, H.A. 'Saudi students' attitudes toward using using social media to support learning'. PhD thesis, King Abdul Aziz University, Jeddah, KSA, 2015.

[16] Alqahtani, S. 'Effects of Social Networking on higher education on Saudi Arabia', in Issa, Isaias, Kommers (eds), Social Networking and Education, Springer, Lecture Notes in Social Networks, 2016, pp.291-304.

[17] Alsurehi, H., & Youbi, A. Towards applying social networking in higher education. International Journal of Academic Research, 2014, 6 (5), 221-229.

[18] Al Lily, A.E. Information Thinness: Saudi Arabia, The Information Society, 2015, 31:5, 407-413, DOI: 10.1080/01972243.2015.1069771.

Real-Time Intelligent Parking Entrance Management

Sofia Belkhala[1], Siham Benhadou[2], Hicham Medromi[3]
Research Foundation for Development and Innovation in Science and Engineering
Engineering research laboratory (LRI), System Architecture Team (EAS)
National and high school of electricity and mechanic (ENSEM) Hassan II University, Casablanca, Morocco[1, 2, 3]

***Abstract*—To help improve the situation of urban transport in the city of Casablanca, we have studied and set up a smart parking system. In this paper, we evaluate the management of the parking entrance utilising artificial intelligence. In addition, we want to establish the limits of our solution and its ability to respond to different requests in real time.**

Keywords—Urban mobility; smart parking; IoT; artificial intelligence; agent; multi agent system; queuing theory

I. Introduction

The world today has become a major project for the city of tomorrow, visions led by investors, idea carriers, and the governance of the city, whose objectives are the modernization of cities while exploiting the emergence of technology in everyday life through connected objects. According to [1], a large part of our life is based on a virtual world, where we have an interconnection between services, products, infrastructure and citizens. This interconnection was made possible thanks to the Internet, today we have what is called: Internet of Things IoT, Internet of People IoP, Internet of energy IoE, and Internet of service IoS.

The emergence of technology is seen as a winning card to solve mobility problems that will contribute to the development of an intelligent city, providing these citizens with an efficient and sustainable transport network. Indeed, the car fleet has grown in recent years, this growth was not accompanied by modernization or improvement of infrastructures. In addition, the starting point and arrival point of a car is the parking, yet there are not enough of them in cities. This lack of parking has caused several problems:

- Unregulated rates
- Air pollution: due to CO2 emissions
- Congestion due to drivers looking for parking spaces
- Waste of time

As has been said, current technologies (Iot, IoP) will be the right choice for efficient, sustainable systems capable of making decisions in real-time. In this perspective, different solutions with different approaches were proposed, either for the collection of information on the Internet of Things, crowdsourcing, predictions for a parking space () or the services offered.

In order to have systems with the above-mentioned characters, the use of multi-agent systems MAS has proven necessary given their ability to operate in distributed environments, with real-time decision-making.

In the state of the art, there are works that have used multi-agent systems to address different parking issues, [2] have used a MAS-based simulation environment to analyze the behavior of motorists looking for parking spaces in the urban environment. Author in [3] made between a Machine to Machine based architecture using MAS for the governance of the solution, according to them the use of MAS will allow them to work under complex conditions in distributed environments.

Author in [4] modeled the driver by an agent who has the behavior to drive, look for a parking space, parking and departure. In addition, what they have deployed includes reactions to different scenarios such as price variance, or lack of places. Authors in [5] established a system of negotiation and guidance based on agents who cooperate and coordinate with each other to negotiate parking prices and calculate the shortest route according to the driver's requirements.

On the other hand, researchers from Tunis [6] have been working on reducing the search time for parking spaces by a multi-agent approach to ensure coordination between drivers and the parking network in the city.

Our team works on a closed, indoor car park offering different services to its users, to ensure a good service, it is necessary to ensure that the time to be served must be minimal therefore the system must be able to manage the inputs/outputs with the different conditions at the time t and deliver real-time responses. Indeed, according to Queue city: Authority and trust in the waiting line, queues have an effect on the emotional geography of the city of boredom, relaxation or even pain hope or rage and existential anxiety in relation to the order of queues and disorder.

In this paper, we will look at what is happening at the entrance, coordination between agents, decision-making and finally a study of the parameters of the queue in the case of our parking. After defining the multi-agent systems, we will introduce the agents responsible for input management, then we will move on to defining the parameters of the queue and then we end with a conclusion.

II. Proposed System

A. Multi Agent System Description

During the phase of the parking tests, we noticed the presence of different elements, a non-linear system, heterogeneous data, different decisions to be taken according to the case, to be able to manage all these challenges and ensure a good functioning of the parking we thought of using artificial intelligence and more precisely the notion of agents.

Over the years, several definitions have been given to agents, but a common point between these definitions is that the agent is an autonomous entity with skills of perception, communication, and action on its environment in order to achieve the goals assigned to it. These agents can form or integrate a group of agents, which is called a multi-agent system. In this system, tasks, resources, and intelligence are distributed; the goal assigned to the system can be achieved through communication between agents. These capabilities will allow us to have an extensible system which will be a good tool to add modules easily. Besides, when an agent integrates a system, it allows him to evolve without changing his internal structure, desires or beliefs.

As we defined before, agents are Autonomous entities with specific goals, therefore the goal is to being able to develop a system that brought together these entities, selfish and adaptable and that cooperate at the same time. During development, we must have a system that brings together different cooperative agents, communicating and jointly planning the actions to be taken without competition between them. In addition, agents try to maximize their gains and this can only be done with the development of their knowledge independently as they go along through cooperation rather than competition. The authors [7] also confirm that multi-agent systems are based on the notion of cooperation, which distinguishes them from other disciplines such as expert systems.

To successfully carry out the cooperation mission, agents must be able to exchange their shared objectives, strategy and plan in order to successfully carry out their mission. Agents must communicate because if they do not communicate, it will be more appropriate to break down the tasks in such a way that it is independent, and can be carried out just by an agent [8]. Indeed, agents need a unified language that allows them to communicate with each other and with their environment, if an agent was faced with solving a problem for which he does not have the necessary knowledge, or if he has incomplete plans, it may be possible to solve the problem by communicating with his environment. To ensure exchange between agents, several approaches have been used from Remote Procedure Call (RPC) or Remote Method Invocation (RMI), to CORBA and Object Request Brokers (ORB's), but Foundation for Intelligent Physical Agent Agent Communication Language FIPA-ACL, the language we have decided to work with, remains the best because it does not only deal with simple objects, and does not describe states as a procedure or method but rather in a declarative language [9], in addition FIPA will allow us to specify how each agent will interact with its environment without touching the agent's internal architecture, which will expand the agent's knowledge and learning domain while keeping its foundations.

At the entrance to the car park, we have installed different components: a barrier, display screens, an OCR camera, an RFID antenna, and a ticket terminal. To guarantee the autonomy and proper functioning of the entrance, we have developed and implemented different agents, in the rest of this paper, we will describe a scenario at the parking entrance to clearly identify the role of each agent. Fig. 1 shows the different blocks.

B. Scenario

When a person arrives at the entrance to the car park, the entry agent must decide whether to initiate the entry process, this decision will depend on the information shared by the control agent about the available spaces, or the reservations made.

Suppose that there are places available, two tasks are triggered in parallel (1) capture of the license plate, and (2) identification of the driver. (1) This is done using the OCR camera, the plate is then sent to the entry agent for verification. For (2) the RFID agent identifies whether the driver at the entrance is a customer or a simple visitor, if he is a customer, the agent determines if his balance is critical, or he has problems with his subscription, so he shares messages with the display agent to inform the driver. The entry agent also sends messages about the customer to preference Agent, who is responsible for identifying the customer's preferences (person with reduced mobility, with an electric car, etc.), as stated by the history of use. If the driver is not a customer, then he must take a ticket, the ticket number is then sent to the entry agent.

After these two processes (identification and registration of license plate), the entry agent, in cooperation with the control agent, authorizes entry, in parallel the preferred agent chooses the best place to be assigned to him so that the display agent takes care of guidance through the screens in the middle of the car park. Table I resumes the description of each agent, his wishes and his triggers.

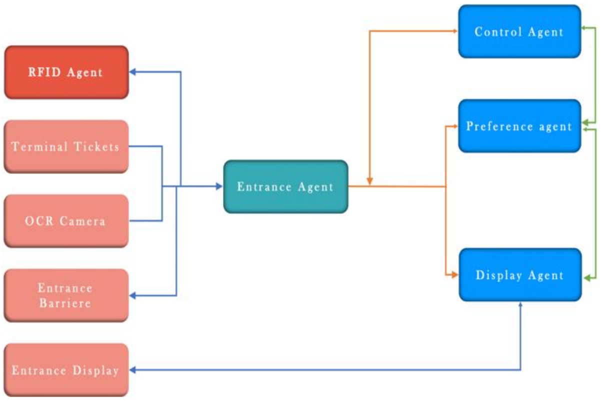


Fig. 1. The Agents in Charge of Entrance Management.

TABLE I. AGENTS DESCRIPTIVE

Agent	**Description**	**Desire**	**Trigger**
RFID Agent	Agent responsible for identifying the type of driver and the car, and verifying the validity of the subscription if it is a customer.	manage drivers at the entrance properly.	Message from the RFID or ticket terminal
Entrace Agent	Agent responsible for entrance management, he is the middleman between the parking area and the entrance	Avoid queues at the entry level Manage entries	Message received from the RFID agent, or control agent
Control Agent	Agent that is in communication with all system agents. It determines, in collaboration with the entry and preference agents, the most appropriate location and then communicate it to the display agent.	Minimize anomalies	Messages from the other agent
Display Agent	he is in charge of ensuring communication with drivers through screens implemented at different levels of the car park. he must determine which message to display (depending on the product cases) and on which screen (those at the entrance, exit...)	Transmit the message correctly to drivers.	Messages from the other agent
Preference Agent	Determines customer preferences based on their usage history.	Satisfying customers	Message from the entry agent

III. Parking Waiting Line

When developing a solution for managing entry, we found that we need to do a study of the queue.

Indeed, the car park is an entity offering services to drivers, by offering these services the phenomenon of queuing can occur either at the entrance or exit. Queue modeling will allow us to test the performance of our system and its ability to manage requests while respecting real time constraints, and find a solution to this phenomenon (Fig. 2).

The service at the entrance follows a First In First Out (FIFO) model, In order to model the queue, we will assume that the arrival of the cars follows a stochastic model such that all events are independent, i.e. the arrival of the Vi car does not depend on the arrival of the Vi+1 car or Vi-1. In addition, it is also assumed that the arrivals follow a parameter fish law and the number of drivers admitted follows an exponential parameter law.

According to Kendal's notation our system follows the following model: M/M/1/ or:

- M: represents the Markovian law associated with arrivals.
- M: represents the Markovian law associated with departures.
- 1: Represents the number of servers (for our case we have only one server)
- represents the length of the tail, which is infinite for our case

For our system we assume that the length of the tail is infinite, therefore we must make sure that the fraction / must be less than 1 because otherwise (i.e. />1), the length of the tail will increase and therefore people who will join the system after may not be served.

On the other hand, during the waiting period, two phenomena can occur.

Balking: when a driver arrives at the entrance but decides to leave without joining the queue, this departure can be forced either if there is no more space available (forced balking) or he estimates that the waiting time will be long for him don he decides to leave (unforced balking)

Reneging: when a driver joins the queue, so he automatically joins the system, but decides to leave halfway.

Our system has no memory, so it has memory less, i. e. during a very small interval only one event can occur, whether it is an arrival or a service.

From this hypothesis, we can deduce that in a time (t+h) only the following events can occur:

- 1 arrival 0 Service
- 0 Arrival 1 service
- 0 Arrival and 0 services

Based on these assumptions, we can deduce the probability P_n : the probability in the stable state u system that there is n person in the system is written as follows:

$$(\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+1} \qquad (1)$$

Based on the basic conditions and events that may occur, we find that P_0 and P_1 are linked by the following formula:

$$\mu P_1 = \lambda P_0 \qquad (2)$$

So for n=1 and replacing equation (1) by the value of equation (2) we find:

$$P_2 = \left(\frac{\lambda}{\mu}\right) P_1 = \left(\frac{\lambda}{\mu}\right)^2 P_0 \qquad (3)$$

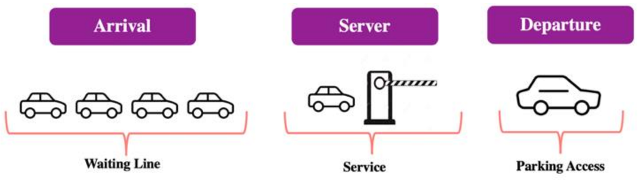


Fig. 2. Queue Model.

So :

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 \quad (4)$$

On the other hand,

$$\sum_{i=1}^{\propto} P_i = 1 \quad (5)$$

We have $\frac{\lambda}{\mu} < 1$, so applying the formula of the geometric sequence we have.

$$P_O + P_1 + P_2 + \cdots \infty = 1 \quad (6)$$

According to (3) :

$$P_0 + \rho P_0 + \rho^2 P_0 + \cdots \infty = 1 \quad (7)$$

$$P_0 \left(\frac{1}{1-\rho}\right) = 1 \quad (8)$$

$$P_0 = 1 - \rho \quad (9)$$

Knowing that $\rho = \frac{\lambda}{\mu}$

So the probability P_n can be written as follows:

$$P_n = \rho^n (1 - \rho) \quad (10)$$

On the other hand, for a good management of the parking we must be interested in 4 parameters:

- Length of the system Ls: excepted number of persons who are at the system including the person who is being served

$$L_s = \frac{\rho}{1-\rho} \quad (11)$$

- Queue length Lq: expected number of people waited for service

$$L_q = L_s - \rho \quad (12)$$

- Waiting time in the system Ws.

$$W_s = \frac{L_s}{\lambda} \quad (13)$$

- Waiting time in the queue Wq.

$$W_q = \frac{L_q}{\lambda} \quad (14)$$

These parameters depend on the probability of the existence of n people in the system.

IV. Conclusion

In this paper, we have focused on the management of entry through an intelligent parking system from a decision-makers viewpoint and collaboration between agents. As well as considering existing research that discusses the phenomenon of the parking queue. In this paper we have used the theory i.e. the model and queue equations to carry out a simulation of the model using data collected from the real car park for study, aiming to improve the communication and collaboration algorithms between the agents of the system.

References

[1] M. Lom, and O. Pribyl, and I. N. Sneddon, “modeling of smart city building blocks using multi-agent systems,” Neural Network World 27(4):317-331 · January 2017.

[2] K. Farkas, and I. Lendák “Simulation Environment for Investigating Crowd-sensing Based Urban Parking,” 2015 Models and Technologies for Intelligent Transportation Systems (MT-ITS), 3-5. June 2015. Budapest, Hungary.

[3] M. Bilal, C. Persson, F. Ramparany, G. Picard, and O. Boissier, “Multi-Agent based governance model for Machine-to-Machine networks in a smart parking management system” 2012 IEEE International Conference on Communications (ICC), 10-15 June 2012, in Ottawa, ON, Canda.

[4] I. Benenson, K. Martens, and S. Birfir, “PARKAGENT: An agent-based model of parking in the city,” Computers, Environment and Urban Systems, Volume 32, Issue 6, November 2008, Pages 431-439.

[5] S. Y. Chou, S. W. Lin , and C. C. Li, “Dynamic parking negotiation and guidance using an agent-based” platform Expert Systems with Applications Volume 35, Issue 3, October 2008, Pages 805-817.

[6] S. Benhassine, R. Harizi , and R. Mraïhi, “Intelligent Parking Management System by Multi-Agent Approach: The case of Urban Area of Tunis” 2014 International Conference on Advanced Logistics and Transport (ICALT), Hammamet, Tunisia, 1-3 May 2014.

[7] J. E. Doran, S. Franklin, N. R. Jennings, And T. J. Norman “On Cooperation In Multi-Agent Systems” The Knowledge Engineering Review Volume 12, Issue 3 , pp. 309-314, September 1997.

[8] L. Panait, and S. Luke “Cooperative Multi-Agent Learning: The State of the Art” Autonomous Agents and Multi-Agent Systems, Volume 11, Issue 3, pp 387–434, November 2005.

[9] Y. Labrou, T. Finin, and Yun Peng “Agent communication languages: the current landscape” IEEE Intelligent Systems and their Applications, Volume: 14 , Issue: 2 , Mar/Apr 1999.

LUCIDAH

Ligative and Unligative Characters in a Dataset for Arabic Handwriting

Yousef Elarian[1]
Teaching and Learning Coaching
Interserve®
Madīnah, Saudi Arabia

Irfan Ahmad[2]
Information and Computer Science Department
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia

Abdelmalek Zidouri[3]
Electrical Engineering Department
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia

Wasfi G. Al-Khatib[4]
Information and Computer Science Department
King Fahd University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia

***Abstract*—Arabic script is inherently cursive, even when machine-printed. When connected to other characters, some Arabic characters may be optionally written in compact aesthetic forms known as ligatures. It is useful to distinguish ligatures from ordinary characters for several applications, especially automatic text recognition. Datasets that do not annotate these ligatures may confuse the recognition system training. Some popular datasets manually annotate ligatures, but no dataset (prior to this work) took ligatures into consideration from the design phase. In this paper, a detailed study of Arabic ligatures and a design for a dataset that considers the representation of ligative and unligative characters are presented. Then, pilot data collection and recognition experiments are conducted on the presented dataset and on another popular dataset of handwritten Arabic words. These experiments show the benefit of annotating ligatures in datasets by reducing error-rates in character recognition tasks.**

Keywords—Arabic ligatures; automatic text recognition; handwriting datasets; Hidden Markov Models

I. Introduction

Arabic script is widely used, not only for the Arabic language, but also for other languages like Urdu, Jawi, and Persian [1]. Despite such importance of the Arabic script, research in technologies that serve it still remains underdeveloped [2]. This is mainly attributed to the scarcity of specialized resources and algorithms that take the peculiarities of the Arabic script into consideration [2] [3].

Among the peculiarities of the Arabic script is that it is inherently (and diversely) cursive. Twenty-two out of the 28 Arabic letters must be connected to the letter that comes next to them in a word. The general form of connection is via short horizontal extensions, shown in the second row of Table I. Some characters, however, also accept to be connected in special compact forms known as *ligatures* [4] (or typographic ligatures [5]). Examples of Arabic ligatures are shown in the last row of Table I.

Most ligatures are optional, i.e., when certain sequences of characters (that we refer to as “ligatives”) occur, they can be either written in the normal horizontal connection form or as a ligature, depending on the writer’s choice.

Many character sequences are not ligative (or “unligative”, as we refer to them); i.e., they do not have optional ligature forms. One particular family of ligatures (“لا” or “ـلا” family, with different Hamza (“ء”) and Madda (“~”) combinations, namely, “لأ”, “ـلأ”, “لآ”, “ـلآ”, “لإ”, “ـلإ”) is considered obligatory, i.e., its constituent LAM and ALEF characters cannot be connected in the common horizontal form. Table I shows different connections of a ligative, an unligative, and an obligatorily ligative examples.

Ligatures are often used in handwriting for their esthetic effects and their compactness. Their presence, however, can add complexities to tasks like Arabic character recognition and synthesis. Researchers in these and similar fields have attempted to handle some ligatures [6] [7] [8] [9], but a comprehensive and systematic list of ligatives was only available after the initial work presented by the authors in [4].

In addition to handwriting, Arabic ligatures are increasingly being incorporated in modern computer fonts to help make texts more compact, beautiful, and as an alternative method for paragraph justification [10] [11].

The "Arabic Presentation Forms-A" of the Unicode standard encodes more than 300 Arabic and Extended-Arabic ligatures [12]. We notice, however, that their set suffers from incompleteness and inconsistencies. For example, the “ئخ” ligature (as in “دائخ”) is absent in Unicode despite the presence of similar ligatures such as “ئج”, “ئح”, “نج”, “نح”, and “نخ” (cf. [4]).

To the best of our knowledge, Arabic ligatures were not systematically analyzed and studied before. In this paper, we present a detailed study of Arabic ligatures that lead to the design of LUCIDAH, a Ligative and Unligative Dataset for Arabic Handwriting. LUCIDAH is especially designed to be representative of both: ligatures and characters in their non-ligatured forms. It is also designed to be practically concise and natural; based on the guidelines in [4].

TABLE. I. EXAMPLES OF (A) A LIGATIVE, (B) AN UNLIGATIVE AND (C) AN OBLIGATORILY-LIGATIVE CHARACTERS

Mode of connection	Ligative	Unligative	Obligatorily
Unconnected Character-shapes	لـ ـمـ ـح	لـ ـس	لـ ـا
Non-Ligature Connections	لمح	لس	NA
Ligature Connections	لمح	NA	ﻻ or ﻻ

The rest of this paper is organized as follows. In Section II, background on Arabic characters in several typographic models is presented. Section III is devoted to the discussion of ligatures and their categorization. In Section IV, the design of LUCIDAH is described. In Section V, implementation issues such as the data collection forms, the demographic distribution of writers, and some scanning parameters are discussed. In Section VI, text recognition experiments are presented to highlight the impact of annotating ligatures on recognition error-rates. Finally, conclusions are presented in Section VII.

II. BACKGROUND ON THE ARABIC SCRIPT

The Arabic script consists of 28 alphabetic letters that are mapped to the 36 computer characters (shown and explained in details in the subsection of the Arabic Typographic Models). Most Arabic characters must connect to subsequent characters to compose words. A few Arabic characters, however, do not connect to subsequent characters (e.g., Letters “ا” and “ر” in Fig. 1). If any of the non-connecting characters occurs before another letter in a word, they cause the word to split into disconnected *Pieces of Arabic Words* (PAWs) [13] [14] [15] or *subwords* [16]. (In addition to that, there is one non-connecting character, “ء”, that does not allow previous characters to connect to it, even when the previous character is a connecting one; hence, its occurrence also causes words to be split into PAWs).

A. Arabic Character-Shapes

Each Arabic character takes a “*character-shape*” based on its position in a PAW, as detailed in the following subsection. Arabic traditionally introduces four *character-shapes*, based on characters' context (i.e. connection with previous and next characters in a PAW). If a PAW consists of only one character, that character takes the *Alone* (A) character-shape. PAWs with more than one letter must start with a *Beginning* (B) character-shape and end with an *Ending* (E) character-shape. If the PAW consists of three or more letters, there would be characters between the Beginning and the Ending character-shapes. These must be *Middle* (M) character-shapes. Table II shows examples of (A), (B), (M), and (E) character-shapes of the letter "س".

Fig. 2 shows an example word consisting of three PAWs and annotates its character-shapes (as Alone (A), Beginning (B), Middle (M), and Ending (E)). Note that Arabic text is written from right to left; hence, the Beginning character-shape of an Arabic multi-lettered PAW appears at the right end of the PAW, whereas the Ending character-shape comes at the leftmost end of it.

Non-connecting letters do not connect, so they do not have Beginning (B) or Middle (M) character-shapes. In addition, there are two characters (“ة” and “ى”) that linguistically can only occur at the end of a word, and hence are treated as non-connecting. Finally, as was mentioned earlier, there is one character, “ء” that does not accept connections from either side, and hence; it only has the Alone (A) character-shape.

B. Arabic Typographic Models

There are several Arabic typographic models. The traditional model introduces up to 4 character-shapes per letter (connecting letters having four shapes, non-connecting letters having two, and “ء” has one). The traditional model encompasses more than a hundred character-shapes, which can be a large number for some applications [17]. Hence, several other models were introduced to represent Arabic script with less numbers of shapes.

One of these reduction models is the 2-Shapes model [18]. The 2-Shapes model clusters the (B) and (M) shapes together and the (A) and (E) shapes together, whenever the differences are merely the connecting extension found preceding the (M) and (E) shape, as seen in Fig. 3 (left and middle parts).

The 1-Shape model introduces *root shapes*, or basic glyph parts that are present in all shapes of a character [17]. In addition to the common root shapes, many (A) and (E) character-shapes also have *tail*s. For example, the character “ص” in Fig. 3 (right) has the “صـ” root in all of its shapes, and the “ں” tail in its (A) and (E) shapes.

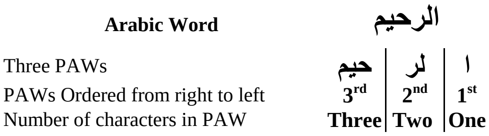


Fig. 1. An Arabic Word Divided into its PAWs.

TABLE. II. DIFFERENT CHARACTER-SHAPES OF THE ARABIC CHARACTER "س" BASED ON ITS CONNECTIONS TO PREVIOUS AND NEXT CHARACTERS

Shape	Example Character-Shape	Example Word	Connected to previous character	Connected to next character	Size of PAW[a]
(A)	س	درس	No	No	1
(B)	سـ	أسف	No	Yes	2+
(M)	ـسـ	تسامح	Yes	Yes	3+
(E)	ـس	خس	Yes	No	2+

[a] The "+" sign after a number *X* in this column denotes "*X* or more characters".

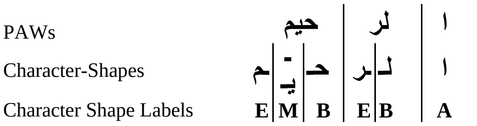
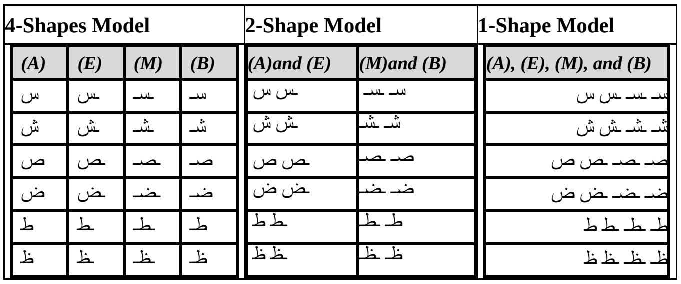


Fig. 2. An Arabic word divided into PAWs with character-shapes annotated.



4-Shapes Model				2-Shape Model		1-Shape Model
(A)	(E)	(M)	(B)	(A)and (E)	(M)and (B)	(A), (E), (M), and (B)
س	ـس	ـسـ	سـ	ـس س	سـ ـسـ	سـ ـسـ ـس س
ش	ـش	ـشـ	شـ	ـش ش	شـ ـشـ	شـ ـشـ ـش ش
ص	ـص	ـصـ	صـ	ـص ص	صـ ـصـ	صـ ـصـ ـص ص
ض	ـض	ـضـ	ضـ	ـض ض	ضـ ـضـ	ضـ ـضـ ـض ض
ط	ـط	ـطـ	طـ	ـط ط	طـ ـطـ	طـ ـطـ ـط ط
ظ	ـظ	ـظـ	ظـ	ـظ ظ	ظـ ـظـ	ظـ ـظـ ـظ ظ

Fig. 3. Examples of Character Groups in the 4-, 2-, and 1-Shape(s) Models.

Appendix I shows Arabic characters in different models. There are few characters that do not share the root shapes across some of their character shapes (such as “ـهـ” and “ك”) . These may not fully cluster their shapes in the 2- and 1-Shape model.

Table III shows examples of a character that would only fit in the 4-Shapes model (since its character-shapes do not share any common glyphs), another that fits in the 2-Shapes model, and a third character that can fit in any model including the 1-Shape model.

TABLE. III. EXAMPLES OF CHARACTER-SHAPES WITH THEIR SMALLEST (BEST) POSSIBLE SHAPE MODELS

Best Model Fit	**(B)**	**(M)**	**(E)**	**(A)**
4-Shapes	هـ	ـهـ	ـه	ه
2-Shapes	كـ	ـكـ	ـك	ك
1-Shape	صـ	ـصـ	ـص	ص

More reduction to the numbers of representative character-shapes of a dataset can be achieved via the *dot-less* (DL) model [19] [20]. The dot-less reduction model clusters Arabic character-shapes with identical main glyphs that differ only in upper or lower dots, upper or lower Hamza "ء", or upper Madda “ٓ”. In Fig. 4, we show the examples of Fig. 3 after DL reduction.

The dot-less model exploits resemblances among different letters whereas the Shapes model exploits resemblances among character-shapes. The two reductions can be combined. The groupings of character-shapes in DL reduction are shown in Appendix I.

The counts of character-shapes for different typographic models are displayed in Table IV. The table does not take into consideration counts of the secondary glyphs (dots, Hamza’s, Madda’s, extensions, and tails) nor the different allographs that some letters may have (e.g., the dent-less س (ٮ)).

4-Shapes DL				**2-Shapes DL**		**1-Shapes DL**
(A)	***(E)***	***(M)***	***(B)***	***(A)and (E)***	***(M)and (B)***	***(A), (E), (M), and (B)***
س ش	ـس ـش	ـسـ ـشـ	سـ شـ	س ش	ـسـ سـ	سـ شـ and the tail ں
ص ض	ـص ـض	ـصـ ـضـ	صـ ضـ	ص ض	ـصـ ضـ	صـ ضـ and the tail ں
ظ	ـط ـظ	ـطـ ـظـ	طـ ظـ	ط ظ	ـطـ ـظـ	ط ظ

Fig. 4. Examples of Character Groups in the 4-, 2-, and 1-Shape(s) DL Model.

TABLE. IV. COUNTS OF CHARACTER-SHAPES FOR DIFFERENT TYPOGRAPHIC MODELS

Model	**(A)**	**(E)**	**(M)**	**(B)**	**Total**
Traditional	36	35	23	23	117
Dot-Less 4-Shapes	19	18	11	11	59
2-Shapes	35 + 5[a]		23 + 3[b]		66
Dot-Less 2-Shapes	18 + 3[c]		11 + 2[d]		34
1-Shape	45				45
Dot-Less 1-Shape	24				24

[a] Extra (A) shapes correspond to ء, ع, غ, ه and ة.

[b] Extra (M) shapes correspond to ـعـ, ـغـ and ـهـ.

[c] Extra (A) shapes correspond to ء, عand ه.

[d] Extra (M) shapes correspond to ـعـ and ـهـ.

In general, counts of glyphs should not be the only factor considered in deciding on a model, since reduction can impose more challenges in dealing with the secondary smaller glyphs.

III. CATEGORIZATION OF LIGATIVES AND LIGATURES

In this section, we attempt to categorize ligatures. First, we notice that there are some Arabic character-shapes that always allow their preceding and connecting characters to form ligatures [4]. We refer to such character-shapes as *Omni-ligative*. Omni-ligatives can be of either the Middle or the Ending shapes of a character (as they must be connected by their preceding characters).

The Omni-ligative character-shapes, using the dot-less model, are: “ـحـ”, “ـح”, “ـمـ”, “ـم”, “ـهـ”, “ـصـ”, “ـص”, “ـطـ”, and “ـط”; the last four being included after revising what was initially reported in [4]. Whether the “ي” character-shape is Omni-ligative or *Partio-ligative* can be debated (Partio-ligative is a term we coined from "partial" to indicate ligatives that are not Omni-ligatives). Here, we consider it a partio-ligative since a short concatenation stroke can be noticed when preceded by some characters making their combination more like an unligative.

The first five Omni-ligatives (viz. “ـحـ”, “ـح”, “ـمـ”, “ـم”, and “ـهـ”) are connected with a stroke that descends when writing (from the right to the left where the Omni-ligative character is to be connected from its top). The four next Omni-ligative characters get ligatured via strokes that ascend diagonally from their bottom left corner. We refer to these as “ascending Omni-ligatives”.

Descending ligatures can be further categorized according to their first-character into ligatures that involve inversion or flipping the first character down and others that connect from the top of a cusp without returning from the cusp to the baseline. The different categories of Omni-ligative connections are demonstrated in Table V.

Ligatures that result from partio-ligatives can also be categorized according to some shared patterns. For example, some involve top of the cusps connections, such as “سر”, “ـسر”, “صر”, “ـصر”, and “٧” (or the v-like allograph of “هـ”). Notice that characters with several allographs can have different ligativity with each allograph. Another example is sticking the circular base of “ک” and “ـک” with a highly ascending character like “ا” or “ل”.

TABLE. V. EXAMPLES OF ASCENDING AND DIFFERENT CATEGORIES OF DESCENDING CONNECTION STROKES IN OMNI-LIGATIVES

Connection Stroke	**Ligatured**	**Not ligatured**
Descending from right	جح	جح
Ascending from left	قط	قط
Descending from right with flipped down first-letter	بح	بح
Descending from the top of a cusp at the right	سح	سح

TABLE. VI. EXAMPLES OF DIFFERENT CATEGORIES OF CONNECTION STROKES IN PARTIO-LIGATIVES

Form	**From top of cusps**					**High link**	**"ک" and "کل"**	**Closed letter**	**Higher cusp**
Not ligatured	سي	سر	حبر	صر	نهر	ين	كا	حد	لبيتنا
Ligatured									

Some calligraphic styles (e.g., [11] [21]) add more restricted rules, but since these are not widely used in everyday handwriting, we do not include them in the ligatives list. The "حـ" and "ـحـ" character-shapes, for example, are sometimes made closed ( ) especially before an ascending character (like "أ", "ـد" or "ـل"). Similarly, some writing styles [21] may encourage making one cusp taller than usual if more than three cusps occur in sequence. Examples of partio-ligative ligature categories are displayed in Table VI.

IV. DATASET DESIGN

Researchers have recognized the role of ligatures in Arabic datasets but lacked resources that systematically cover them. Research in text recognition [6], including popular Arabic handwriting datasets [7] [8], attempted to manually annotate ligatures. The absence of a comprehensive list of ligatures prior to [4] made recognizing ligatures difficult. In this section, we follow the guidelines and lists from [4] to design representative yet concise ligative and an unligative datasets.

Corpora and dataset design consider obtaining concise yet representative samples to reflect certain characteristics of a language. Representativeness in corpus and dataset design is defined as "the extent to which a sample includes the full range of variability in a population [22]". Sampling, when used in the context of corpora, refers to the process of selecting limited-sized yet representative texts that are expected to reproduce the characteristics of the underlying language. The sampling theory addresses important concerns such as the sampling unit and the sampling frame [23]. Moreover, we introduce the term *representing unit* to refer to the smallest writing unit that would assure representativeness and conciseness. The representing unit can be one (or more) word(s), a character, a character-shape, or even a pen stroke. By *representation criteria*, we refer to the set of units that need to be present in the dataset for it to be considered representative. The representation criteria can be, for example, to cover character-shapes under a certain reduction model (see Section 2). Representativeness under some reduction model can be assured either for each writer or collectively for sets of writers. We refer to the number of writers ensuring representation under some criteria as *representative forms.*

LUCIDAH design is mainly concerned with selecting adequate representing units, representation criteria, and representative forms to achieve an acceptable level of representativeness with concise and naturally written units. Typically, representativeness and conciseness; conciseness and naturalness; naturalness and representativeness may have conflicting requirements. In the presented dataset design, we attempt to balance all factors to reasonable levels via the exploitation of the powers of different reduction models for different parts of the dataset.

LUCIDAH has two main parts: the ligative and the unligative parts. For the representativeness of undistorted character-shapes and ligatures, LUCIDAH must contain, not only all the desired ligatures, but also all the desired character-shapes in unligative forms [4]. These parts are addressed by careful selection of texts containing characters with extra-care to select their preceding and following characters. The design of the ligative and unligative parts may have different requirements; and hence, each of these parts is discussed in a separate subsection, below.

A. Design of the Ligative Part

Ligatives, in general, can involve more than two characters, if one or more consecutive Middle character-shapes keep forming ligatives with their next characters. Sequences of characters are frequently referred to as *n*-grams. The problem of *n*-grams is that the counts of their combinations increase exponentially with the sizes of their constituent characters, *n*. This clearly makes collecting *n*-grams intractable. We decided to treat *n*-gram ligatives as (n - *1*) connecting bigrams. Hence, we need to limit the representing unit of the ligative part of LUCIDAH to be within the bigram options only while hoping that bigrams would represent all *n*-grams with an acceptable level of naturalness.

A ligature may only occur if a character connects to its subsequent. Hence, bigram ligatures can be formed when either a B or an M character-shape connects to an M or an E character-shape. The four resulting combinations of connected bigrams (shown with examples in Table VII) are: B and E (denoted as *BE*, and also denotable as *A-bigram*; since it, as a whole, cannot be connected to neither previous nor next characters), B and M (denoted as *BM* and also denotable as *B-bigram*; since it cannot connect to a previous character but connects to a following character), M and M (denoted as *MM* and also denotable as *M-bigram*; since it has to be connected to both a previous and a next character), and M and E (denoted as *ME* and also denotable as *E-bigram*; since the whole ligature is not connected to a next but to a previous character).

Similarly, connected trigram can be expressed as BMM or a B-trigram, BME or an A-trigram, MMM or a M-trigram, and MME or an E-trigram. We counted the possible bigram combinations and bigram ligatives (including the "ascending Omni-ligatives") and these are displayed in Table VIII.

TABLE. VII. EXAMPLES OF A STANDALONE LIGATURE AND LIGATURES AT THE BEGINNING, MIDDLE, AND END, OF THEIR CORRESPONDING SUBWORD

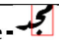
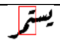
2nd **1st**	**(M)**	**(E)**
(B)	**(BM) or B-Ligative** 	**(BE) or A-Ligative**
(M)	**(MM) or M-Ligative** 	**(ME) or E-Ligative**

TABLE VIII. COUNTS OF LIGATIVE BIGRAMS WITH THE TOTAL NUMBER OF BIGRAMS COMBINATIONS FOR THE DIFFERENT TYPOGRAPHIC MODELS

Model	(BE)	(ME)	(MM)	(BM)	Total
Traditional	247 / **782**	255 / **782**	213 / **529**	212 / **529**	927 / **2622**
Dot-Less (DL)	56 / **198**	56 / **198**	58 / **121**	57 / **121**	227 / **638**
2-Shapes	281 / **884**		240 / **598**		521 / **1482**
(DL) 2-Shapes	65 / **234**		68 / **143**		133 / **377**
1-Shape	291 / **907**				291 / **907**
(DL) 1-Shape	70 / **245**				70 / **245**

Upon inspecting the counts of ligative bigrams under several models, we found that for a tractable size of a ligative part form, we should base the design on the minimum size possible, i.e. the combined dot-less with 1-Shape model. We chose the BE (or A-Ligatures) as representatives. The BE or A-Ligatures bigrams are standalone bigrams that do not come connected to other characters. Hence, they may be written alone with higher levels of naturalness than any of the other types. However, there are some ligatives that do not have standalone forms. These are inserted into short words in the dataset design using the 1-Shape Dot-less model. In addition, we prepared 12 different ligative forms and diversified the dot-less representatives in these so that each set of 12 forms collectively has a *representation criteria* of the 1-Shape model.

To summarize this part, the design of ligative part of the dataset uses standalone ligatures and short words as the representing unit. Each form alone represents the characters under the dot-less and 1-Shape combined criteria. Every 12 consecutive forms, however, can together achieve representativeness under the 1-Shape criteria (without the dot-less reduction).

B. Design of the Unligative Part

Pangrams are texts that contain all characters of an alphabet and lipograms are writings constrained to avoid certain sets of characters [24]. A representative unligative dataset should cover all character-shapes while avoiding ligatives. So, in a sense, the representativeness problem of the unligative dataset can be formulated as a search for pangram of all character-shapes with lipogram conditions to avoid ligatives.

The counts of necessary characters to satisfy the above conditions are significantly smaller than their ligative counterpart, as the bounds in [4] show. Hence, we could afford to use "sentences" as representing units, the "traditional model" as the representation criteria, and "each form" as a representative. These choices allowed us to increase the level of naturalness in the unligative part while maintaining it reasonably concise.

Under the above conditions, we aimed at finding a minimal set [25] of representative sentences for the unligative dataset. Several character-shape pangrams were created. However, the set of sentences in Fig. 5(a), along with the set of A-character-shapes in Fig. 5(b), were chosen for being the most concise. The separation of the eight character-shapes of Fig. 5(b) reduces the total number of required words by eight, since these shapes can only appear once in a word.

The text contains 43 words with 163 character-shapes, 6 of them combined into three instances of obligatory ligatures. Unfortunately, the two conditions of the unligative text cannot be fully fulfilled due to the need of the inclusion of Omni-ligatives, which cannot be guaranteed not to be written as ligatures by any sequence. Hence, we could only aim at avoiding partio-ligatives.

بلغ حاج أن أخاه ظمآن بوادي عوف طفق يسعى لإحضار ثلاث قرب زمزم تنجيه مع سطوع
وهيج الشمس حث عوض الشيخ نوح بصدد ذلك فأكرمه وصب وتكلف وقال للآت أعظم
ضبط سهيل وأشخاص لص الحي. غش راجح غثامة لذا جن بغيظ وانقض انتهت

(a)

خ س ش ط ظ غ ق ك

(b)

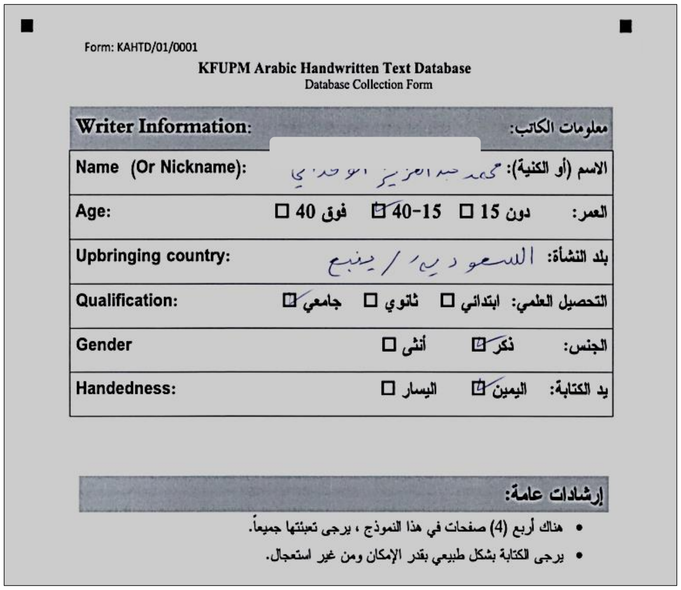
Fig. 5. (a) The Selected text for Collecting the Character-Shapes, and (b) The Complementary Set of Alone Character-Shapes.

V. DATASET IMPLEMENTATION AND COLLECTION

We have designed forms with ligative and unligative parts to collect handwritten samples. Every form is intended to be filled by a distinct and unique writer. The forms were printed, distributed, collected back and analyzed. After discarding incomplete forms, 450 of the collected forms were selected to be scanned at 300 DPI into TIFF colored images.

Parts of the ligative and unligative forms are shown in the figures below. Fig. 6 shows the part where the writers' information is acquired. Fig. 7 depicts the ligative parts grid, with the filling spaces being of equal and vast areas. Each set of 12 of these forms are similar but not identical as discussed earlier. Fig. 8 shows an example of the unligative paragraph along with the corresponding isolated characters.

On all form pages, three aligned filled squares are printed to ease automatic skew detection and correction. The boxes are printed in positions such that their centers of gravity form a right angle with the grids surrounding the content parts of the form. The corner that does not have an aligning box varies depending on the page-number within the form.



Form: KAHTD/01/0001

KFUPM Arabic Handwritten Text Database
Database Collection Form

Writer Information:		معلومات الكاتب:
Name (Or Nickname):	محمد عبدالعزيز الوحداني	الاسم (أو الكنية):
Age:	☐ فوق 40 ☑ 40-15 ☐ دون 15	العمر:
Upbringing country:	السعودية / ينبع	بلد النشأة:
Qualification:	☑ جامعي ☐ ثانوي ☐ ابتدائي	التحصيل العلمي:
Gender	☐ أنثى ☑ ذكر	الجنس:
Handedness:	☐ اليسار ☑ اليمين	يد الكتابة:

إرشادات عامة:

- هناك أربع (4) صفحات في هذا النموذج ، يرجى تعبئتها جميعاً.
- يرجى الكتابة بشكل طبيعي بقدر الإمكان ومن غير استعجال.

Fig. 6. A Scanned Sample of the Writer Information Collection form.

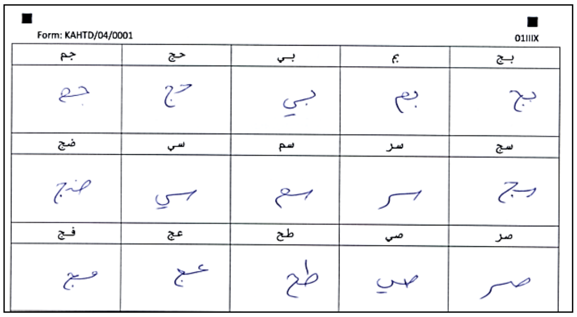
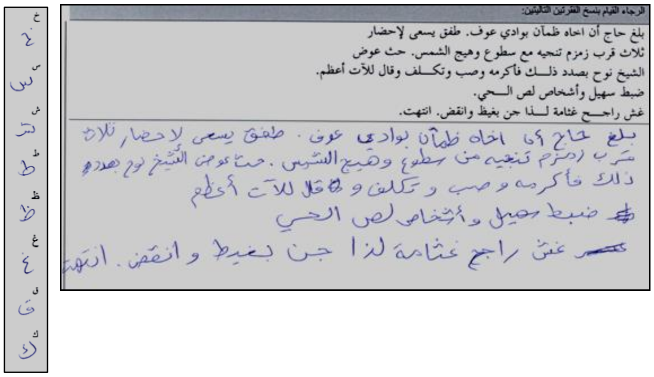


Fig. 7. Two Scanned forms of the Ligative Part of the Dataset.


الرجاء القيام بنسخ الفقرتين التاليتين:
بلغ حاج أن اخاه ظمآن بوادي عوف. طفق يسعى لإحضار
ثلاث قرب زمزم تنجيه مع سطوع وهيج الشمس. حث عوض
الشيخ نوح بصدد ذلـــك فأكرمه وصب وتكـــلف وقال للآت أعظم.
ضبط سهيل وأشخاص لص الـــحي.
غش راجـــح غثامة لـــذا حن بغيظ وانقض. انتهت.

(a) (b)

Fig. 8. A Scanned Sample of Filled (a) isolated Characters and (b) unligative Parts of a form.

Table IX displays a summary of our ultimate design choices related to the unit-selection and the coverage criteria of the different forms of the dataset. The representation criteria looked at the full range from the most restricted reduction-model (the combined dot-less 1-Shape model) to the traditional model. The representative form ranged from half a page paragraph to 12 pages to be filled by 12 writers. The representing units covered isolated characters, Standalone ligatures and words, and even full sentences and paragraphs.

In Table X, we show some statistical information on the regions, genders, writing-hands, and qualifications of the writers. We consider the following three regions: The Arab Peninsula: containing the Gulf countries, Yemen and Iraq, North Africa: containing Egypt, Sudan, and the countries of Northwest Africa, and Levant: containing Syria, Jordan, Palestine, and Lebanon.

TABLE. IX. UNITS AND COVERAGE CRITERIA OF THE DIFFERENT FORMS OF THE DATASET

Parts	**Representing Unit**	**Representation Criteria**	**Representative Forms**
Ligatures	Standalone ligatures and words	Combined dot-less and 1-Shape	1 form
		1-Shape ligatures	12 forms
Unligative text	Sentences	Traditional (all character-shapes)	1 form
Isolated characters	Isolated Characters		

TABLE. X. NUMBERS OF WRITERS IN THE COLLECTED DATASET PER REGION, GENDER, HANDEDNESS AND QUALIFICATIONS

Region	Arab Peninsula	417
	North Africa	22
	Levant	11
Gender	Male	398
	Female	52
Writing Hand	Right Hand	416
	Left Hand	34
Qualification	Intermediate School	4
	High School	386
	B.Sc. / BA	53
	M.Sc. / Ph.D.	7

VI. Experiments and Results

In this section, we will present the text recognition experiments conducted to show the benefit of incorporating the knowledge of ligatures in dataset design for Arabic. First, we present the datasets used for the text recognition experiments. Then, we describe the text recognition tasks, the measures used for reporting the results, and the used text recognition system. This is followed by the description and discussion of the obtained results.

A. LUCIDAH Dataset

We used a part of the LUCIDAH dataset for training recognition. We selected 10 word images from the dataset which contain Omni-ligatives. Some writers wrote these character pairs as ligatures while others did not. A total of 594 word images were used for training and evaluation. Fig. 9 shows sample words from the LUCIDAH dataset.

B. IFN/ENIT Dataset

The IFN/ENIT dataset [8] consists of handwritten images of Tunisian towns and is one of the most popular publicly available datasets for Arabic text recognition research. The original dataset consists of 32,492 images divided into 5 sets (Sets *a*, *b*, *c*, *d*, and *e*). Later, 10,244 more images were added as Sets *f* and *s*. Fig. 10 shows some sample images from the IFN/ENIT dataset. For the experiments presented here, we will use the standard training-test partitions as reported in the literature (e.g., [26] [27]).

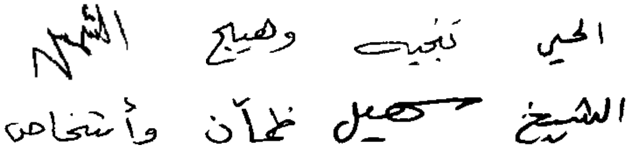


Fig. 9. Sample Text Images from the LUCIDAH Dataset.

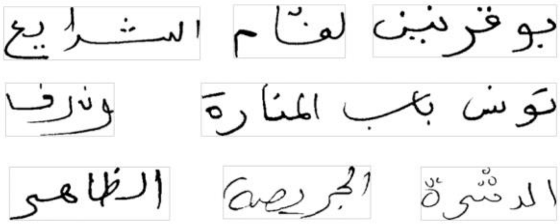


Fig. 10. Sample Text Images from the IFN/ENIT Dataset.

C. Text Recognition Task

We performed character-based recognition after training the system on the text images and two sets of ground-truths: one modeling ligatures and another not modeling them. We used character-based recognition without word dictionaries or language models to adequately show the effects of different modeling choices (ligatures as models versus models with no ligatures). Thus, for the text images, we predict the character sequence using only the character models training without using any n-grams or dictionaries. Accordingly, we report the result in terms of character error-rates (CER):

$$CER\ (\%)\ = \frac{S + I + D}{N} \times 100$$

where;

S is the error due to character substitution,

D is the error due to character deletion,

I is the error due to character insertion, and

N is the total number of characters in the evaluation set.

In addition to the CER, we also report the statistical significance of the recognition results. Statistical test for the difference of two proportions as presented in [28] is used to report the significance interval of the results at 95% confidence level. This helps us compare the results of two systems with high confidence.

D. Recognition System Description

In this section, we present the details of the system used for training and recognizing the word images. The present system is based on Hidden Markov Models (HMMs). We use the HTK tools [29] to build the recognition system. We, first, preprocess the text images by normalizing the height of the images to 96 pixels while keeping the aspect ratio unchanged for each image. This is followed by the feature extraction stage where sliding windows (of 8 pixels wide and the 96 pixels of height and overlap of 4 pixels) were used. Nine features, that are adapted from the work of Wienecke et al. [30] are extracted from the image segment under the window. In addition, we append 9 more derivative features for each window along the horizontal axis. Thus, the feature vector is of size 18.

To train the system, we build a separate model for each character-shape. A character-shape model consists of a continuous HMM with Bakis topology, thereby, allowing the possibility of skipping the consecutive state during state transitions. Hyper-parameters, like the number of states per model and the number of mixtures per states, were decided based on the performance of the recognizer on the development set.

A multi-step approach was followed for model initialization and training. As a first step, the models were initialized by the flat-start (uniform initialization) approach followed by iterative Baum-Welch training. Then, forced alignment was performed on the training data. The forced aligned data was, in turn, used to reinitialize the models using the Viterbi algorithm. Then, iterations of Baum-Welch training were performed on the reinitialized models. Finally, character recognition was performed using the Viterbi algorithm.

E. Text Recognition on LUCIDAH Dataset

In this section, we will present the experiments and results on the LUCIDAH dataset. As the used part of the dataset for the pilot study is small, the training and evaluation, we performed uniform initialization followed by few iterations of Baum-Welch training instead of the previously described multi-stage training.

The first experiments were conducted without models for ligature. Each character-shape was treated as a model. The training set contained 36 different character-shapes modelled into 36 HMMs. A total of 2,696 character-shape instances were available in the training set.

The second set of experiments was conducted utilizing the ligature models. The pilot dataset was manually investigated for the presence of ligatures. The ligatures were treated as separate models in this set of experiments. The training set encompassed 47 models (36 character-shapes and 11 ligature models). Table XI presents the key characteristics of the training set.

Table XII displays the character-based text recognition results for the two systems in terms of character error-rates (CER). Also, statistical significance at 95% confidence is reported. We can clearly see from the table that the system with the ligature models has significantly lower error-rates as compared to the system without ligature models, for a 95% confidence level. The results were encouraging, but since the pilot study only involved a small dataset, we decided to experiment with the IFN/ENIT dataset, too, as we discuss it in the following section.

F. Text Recognition using IFN/ENIT Dataset

In this section, we present the details of the experiments and the results obtained for character-based text recognition on the IFN/ENIT dataset. We used the standard *train-test* configurations as reported in the literature. System hyper-parameters were calibrated based on an evaluation set (Set *d*) with training sets *a, b, and c* (*i.e.* the *abc-d* configuration).

TABLE. XI. KEY MODEL STATISTICS ON LUCIDAH DATASET

	Number of HMMs	**Average number of samples per model**	**Median number of samples per model**
No Ligature Models	36	75	54
Ligature Models	**47**	**53**	**43**

TABLE. XII. CHARACTER RECOGNITION RATES ON THE LUCIDAH DATASET

System Description	**CER (%)**	**Statistical Significance**
System with no ligature models (baseline)	25.04	±1.40
System with ligature models	**21.52**	

TABLE. XIII. SUMMARY OF CHARACTER-BASED TEXT RECOGNITION RESULTS ON THE IFN/ENIT DATASET

System Description	**CERs (%)**			
	Train–Test Configurations			
	abc–d	**abcd–e**	**abcde–f**	**abcde–s**
System with no ligature models (baseline)	40.65	49.68	46.33	55.12
System with ligature models	**37.49**	**44.81**	**41.69**	**51.64**

Table XIII presents the summary of the text recognition results for the IFN/ENIT dataset using the system without ligatures (second row) and the system with ligatures (bottom row). Significance interval of the errors is ±0.35, ±0.38, ±0.32, and ±0.75 for evaluation sets *d*, *e*, *f*, and *s* respectively at 95% confidence level. We can see from the table that the system having ligature models outperforms the system having no ligature models.

Text recognition results on both datasets confirm that annotating ligatures in handwritten Arabic can enhance their recognition performance. Hence, considering ligatures in dataset design is important for text recognition research. The knowledge of the ligatures can help design datasets which can enable collection of ligature samples from the writers, in addition to its other applications.

VII. CONCLUSIONS

The Arabic script is widely used around the world. Arabic has an inherently cursive script where some character sequences can be replaced by more compact forms called ligatures. If ligatures are not distinctly annotated in datasets, their special forms may cause confusions for automatic text recognition systems. To ease the annotation of ligatures, we design and implement a ligative and unligative dataset for Arabic handwriting, LUCIDAH. Several design decisions taken to balance the representativeness, conciseness, and naturalness requirements of the dataset were presented in this paper. Pilot text recognition experiments were conducted on LUCIDAH and IFN/ENIT to show the significant benefits of annotating ligatures in reducing character recognition errors.

ACKNOWLEDGMENT

The authors thank King Fahd University of Petroleum and Minerals (KFUPM) and Interserve® for their support.

REFERENCES

[1] "Britannica Encyclopedia, Arabic Alphabet," [Online]. Available: https://www.britannica.com/topic/Arabic-alphabet. [Accessed July 2019].

[2] M. Shatnawi and S. Abdallah, "Improving Handwritten Arabic Character Recognition by Modeling Human Handwriting," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 15, p. Artivle 3, November 2015.

[3] A. Al-Sallab, R. Baly, H. Hazem, S. B. Khaled, W. El-Hajj and G. Badaro, "AROMA: A Recursive Deep Learning Model for Opinion Mining in Arabic as a Low Resource Language," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 16.4, no. 25, 2017.

[4] Y. Elarian, I. Ahmad, S. Awaida, W. Al-Khatib and A. Zidouri, "Arabic ligatures: Analysis and application in text recognition," in Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, 2015.

[5] Wikipedia, "Typographic_Ligature," [Online]. Available: http://en.wikipedia.org/wiki/Typographic_ligature. [Accessed July 2019].

[6] J. Trenkle, A. Gillies, E. Erlandson, S. Schlosser and S. Cavin, "Advnces in Arabic Text Recognition," in in Symposium on Document Image Understanding Technology (SDIUT 2001), Columbia, Maryland, 2001.

[7] S. Schlosser, ERIM Arabic document database, Environmental Research Institue.

[8] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze and H. Amiri, "IFN/ENIT - Database Of Handwritten Arabic Words," in CIFED, 2002.

[9] Y. Elarian, I. Ahmad, S. Awaida, W. Al-Khatib and A. Zidouri, "An Arabic Handwriting Synthesis System," Pattern Recognition, vol. 48, no. 3, pp. 849-861, 2015a.

[10] M. Elyaakoubi and A. Lazrek, "Justify Just or Just Justify," The Journal of Electronic Publishing, vol. 13, no. 1, 2010.

[11] M. I. Salama, "For the Aesthetics of Arabic Calligraphy," in International Conference for the Aestheitcs of Arabic Calligraphy, Alexandria, 2006.

[12] Unicode Consortium, "Arabic Presentation Forms-A," [Online]. Available: http://www.unicode.org/charts/PDF/U0600.pdf. [Accessed July 2019].

[13] Y. Elarian, A. Zidouri and W. Al-Khatib, "Ground-truth and Metric for the Evaluation of Arabic Handwritten Character Segmentation," in International Conference on Frontiers in Handwriting Recognition, Crete, 2014.

[14] I. Ahmad and G. Fink A., "Class-Based Contextual Modeling for Handwritten Arabic Text Recognition," in Frontiers in Handwriting Recognition (ICFHR), Shenzhen, 2016.

[15] Y. M. Alginahi, "A survey on Arabic Character Segmentation," International Journal on Character Analysis and Recognition (IJDAR), vol. 16, no. 2, pp. 105-126, 2013.

[16] M. Maliki, N. Al-Jawad and S. Jassim, "Sub-word based Arabic handwriting analysis for writer identification," in Proc. SPIE 8755, Mobile Multimedia/Image Processing, Security, and Applications , 2013.

[17] F. Menasri, N. Vincent, E. Augustin and M. Cheriet, "Shape-based Alphabet for Off-line Arabic Handwriting Recognition," in Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on (Volume:2), Parana, 2007.

[18] Y. Haralambous, "The traditional Arabic typecase extended to the Unicode set of glyphs," Electronic Publishing -- Origination, Dissemination and Design (EP-odd), vol. VOL. 8(2 & 3), no. June & September 1995, p. 111–123, 1995.

[19] I. Ahmad and G. A. Fink, "Multi-stage HMM based Arabic text recognition with rescoring," in 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, 2015.

[20] Y. Elarian and F. Idris, "A lexicon of Connected Components for Arabic Optical Text Recognition," in First International Workshop on Frontiers in Arabic Handwriting Recognition, Istanbul, 2010.

[21] M. Hashem, Artist, The Fundamentals of Arabic Calligraphy: a manuscript collecting the different calligraphy styles of Arabic الـ خطاط, الـ عربـ ية طوطالـخ لأذ واع خط ية مجموعة الـ عربـ ي الـ خط قواعد محمد هاشم. [Art]. Calligraphy, 1986.

[22] D. Biber, "Representativeness in Corpus Design," Literary and Linguistic Computing, vol. 8, no. 4, pp. 243-257, 1993.

[23] M. Tony, R. Xiao and Y. Tono, "Unit 2 Representativeness, balance and sampling," in Corpus-Based Language Studies, Routledge, 2006.

[24] D. D. Johnson, B. von Hoff Johnson and K. Schlichting, Logology: Word and language play, Guildford Press, 2004.

[25] H. A. Al-Muhtaseb, S. A. Mahmoud and a. R. S. Qahwaji, "A Novel Minimal Script for Arabic Text Recognition Databases and Benchmarks," International Journal of Circuits, Systems and Signal Processing, pp. 145-153, 2009.

[26] V. Margner and H. E. Abed, "Arabic Hnadwriting Recognitoin Competition," in Frontiers in Handwriting Recognition, 2010.

[27] H. El Abed, M. Kherallah, V. Margner and A. M. Alimi, "Online Arabic Handwriting Competition," International Journal on Document Analysis and Recognition (IJDAR), vol. 14(1), pp. 15-23, 2011.

[28] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural computation , vol. 10.7, pp. 1895-1923, 1998.

[29] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, The HTK Book (for HTK Version 3.2. 1), Cambridge University Engineering Department, 2002.

[30] M. Wienecke, G. A. Fink, and G. Sagerer, "Toward automatic video-based whiteboard reading," International Journal of Document Analysis and Recognition , Vols. 7.2-3 , pp. 188-200, 2005.

APPENDIX

Appendix I: Groups of Arabic characters in the 4-, 2-, and 1-Shapes Normal (Fig. 11) and Dot-Less (Fig. 12) models. Highlighted character groups cannot be merged due to differences between their (B) and (M) and/or between their (A) and (E) shapes.

4-shapes model

(A)	*(E)*	*(M)*	*(B)*
ء			
ا	ـا		
أ	ـأ		
إ	ـإ		
آ	ـآ		
ب	ـب	ـبـ	بـ
ت	ـت	ـتـ	تـ
ث	ـث	ـثـ	ثـ
ج	ـج	ـجـ	جـ
ح	ـح	ـحـ	حـ
خ	ـخ	ـخـ	خـ
د	ـد		
ذ	ـذ		
ر	ـر		
ز	ـز		
س	ـس	ـسـ	سـ
ش	ـش	ـشـ	شـ
ص	ـص	ـصـ	صـ
ض	ـض	ـضـ	ضـ
ط	ـط	ـطـ	طـ
ظ	ـظ	ـظـ	ظـ
ع	ـع	ـعـ	عـ
غ	ـغ	ـغـ	غـ
ف	ـف	ـفـ	فـ
ق	ـق	ـقـ	قـ
ك	ـك	ـكـ	كـ
ل	ـل	ـلـ	لـ
م	ـم	ـمـ	مـ
ن	ـن	ـنـ	نـ
ه	ـه	ـهـ	هـ
ة	ـة		
و	ـو		
ؤ	ـؤ		
ي	ـي	ـيـ	يـ
ى	ـى		
ئ	ـئ	ـئـ	ئـ

2-shapes model

(A)	*(E)*	*(M)*	*(B)*
ء			
ا ـا			
أ ـأ			
إ ـإ			
آ ـآ			
ب ـب		ـبـ بـ	
ت ـت		ـتـ تـ	
ث ـث		ـثـ ثـ	
ج ـج		ـجـ جـ	
ح ـح		ـحـ حـ	
خ ـخ		ـخـ خـ	
د ـد			
ذ ـذ			
ر ـر			
ز ـز			
س ـس		ـسـ سـ	
ش ـش		ـشـ شـ	
ص ـص		ـصـ صـ	
ض ـض		ـضـ ضـ	
ط ـط		ـطـ طـ	
ظ ـظ		ـظـ ظـ	
ع	ـع ـعـ		عـ
غ	ـغ ـغـ		غـ
ف ـف		ـفـ فـ	
ق ـق		ـقـ قـ	
ك ـك		ـكـ كـ	
ل ـل		ـلـ لـ	
م ـم		ـمـ مـ	
ن ـن		ـنـ نـ	
ه	ـه	ـهـ	هـ
ة	ـة		
و ـو			
ؤ ـؤ			
ي ـي		ـيـ يـ	
ى ـى			
ئ ـئ		ـئـ ئـ	

1-shape model

(A)	*(E)*	*(M)*	*(B)*
ء			
ا ـا			
أ ـأ			
إ ـإ			
آ ـآ			
ب ـب ـبـ بـ			
ت ـت ـتـ تـ			
ث ـث ـثـ ثـ			
ج ـج ـجـ جـ			
ح ـح ـحـ حـ			
خ ـخ ـخـ خـ			
د ـد			
ذ ـذ			
ر ـر			
ز ـز			
س ـس ـسـ سـ			
ش ـش ـشـ شـ			
ص ـص ـصـ صـ			
ض ـض ـضـ ضـ			
ط ـط ـطـ طـ			
ظ ـظ ـظـ ظـ			
ع	ـع ـعـ		عـ
غ	ـغ ـغـ		غـ
ف ـف ـفـ فـ			
ق ـق ـقـ قـ			
ك ـك		ـكـ كـ	
ل ـل ـلـ لـ			
م ـم ـمـ مـ			
ن ـن ـنـ نـ			
ه	ـه	ـهـ	هـ
ة	ـة		
و ـو			
ؤ ـؤ			
ي ـي		ـيـ يـ	
ى ـى			
ئ ـئ		ـئـ ئـ	

Fig. 11. Character Groups in the 4-, 2-, and 1-Shape(s) Models.

4-Shape Dot-less Model			
(A)	***(E)***	***(M)***	***(B)***
ء			
ا أ إ آ	ـا ـأ ـإ ـآ		
ب ت ث	ـب ـت ـث	ـبـ ـتـ ـثـ ـنـ ـيـ ـئـ	بـ تـ ثـ نـ يـ ئـ
ن	ـن		
ي ئ ى	ـي ـئ ـى		
ج ح خ	ـج ـح ـخ	ـجـ ـحـ ـخـ	جـ حـ خـ
د ذ	ـد ـذ		
ر ز	ـر ـز		
س ش	ـس ـش	ـسـ ـشـ	سـ شـ
ص ض	ـص ـض	ـصـ ـضـ	صـ ضـ
ط ظ	ـط ـظ	ـطـ ـظـ	طـ ظـ
ع غ	ـع ـغ	ـعـ ـغـ	عـ غـ
ف	ـف	ـفـ ـقـ	فـ قـ
ق	ـق		
ك	ـك	ـكـ	كـ
ل	ـل	ـلـ	لـ
م	ـم	ـمـ	مـ
ه ة	ـه ـة	ـهـ	هـ
و ؤ	ـو ـؤ		

2-Shapes Dot-less Model			
(A)	***(E)***	***(M)***	***(B)***
ء			
ـا ا ـأ أ ـإ إ ـآ آ			
ـب ب ـت ت ـث ث		بـ ـبـ تـ ـتـ ثـ ـثـ نـ ـنـ يـ ـيـ ئـ ـئـ	
ـن ن			
ـي ي ـئ ئ ـى ى			
ـج ج ـح ح ـخ خ		جـ ـجـ حـ ـحـ خـ ـخـ	
ـد د ـذ ذ			
ـر ر ـز ز			
ـس س ـش ش		سـ ـسـ شـ ـشـ	
ـص ص ـض ض		صـ ـصـ ضـ ـضـ	
ـط ط ـظ ظ		طـ ـطـ ظـ ـظـ	
ع غ	ـع ـغ	ـعـ ـغـ	عـ غـ
ـف ف		فـ ـفـ قـ ـقـ	
ـق ق			
ـك ك		كـ ـكـ	
ـل ل		لـ ـلـ	
ـم م		مـ ـمـ	
ه ة	ـه ـة	ـهـ	هـ
و ـو ؤ ـؤ			

hapes Dot-less Model			
(A)	***(E)***	***(M)***	***(B)***
ء			
ا ـا أ ـأ إ ـإ آ ـآ			
ب ـب ت ـت ث ـث		بـ ـبـ تـ ـتـ ثـ ـثـ نـ ـنـ يـ ـيـ ئـ ـئـ	
ن ـن			
ي ـي ئ ـئ ى ـى			
جـ ـجـ حـ ـحـ خـ ـخـ ج ـج ح ـح خ ـخ			
د ـد ذ ـذ			
ر ـر ز ـز			
سـ ـسـ شـ ـشـ س ـس ش ـش			
صـ ـصـ ضـ ـضـ ص ـص ض ـض			
طـ ـطـ ظـ ـظـ ط ـط ظ ـظ			
ع غ	ـغـ ـغ ـعـ ـع		عـ غـ
ق ـق ف ـف		فـ ـفـ قـ ـقـ	
ك ـك		كـ ـكـ	
ل ـل		لـ ـلـ	
م ـم		مـ ـمـ	
ه ة	ـه ـة	ـهـ	هـ
و ـو ؤ ـؤ			

Fig. 12. Character Groups in the 4-, 2-, and 1-Shape(s) Dotless Models.

An Efficient Normalized Restricted Boltzmann Machine for Solving Multiclass Classification Problems

Muhammad Aamir[1], Fazli Wahid[3], Hairulnizam Mahdin[4]
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn, Malaysia

Nazri Mohd Nawi[2]
Soft Computing and Data Mining Centre (SMC)
Universiti Tun Hussein Onn, Malaysia

***Abstract*—Multiclass classification based on unlabeled images using computer vision and image processing is currently an important issue. In this research, we focused on the phenomena of constructing high-level features detector for class-driven unlabeled data. We proposed a normalized restricted Boltzmann machine (NRBM) to form a robust network model. The proposed NRBM is developed to achieve the goal of dimensionality reduction and provide better feature extraction with enhancement in learning more appropriate features of the data. For increment in learning convergence rate and reduction in complexity of the NRBM, we add Polyak Averaging method when training update parameters. We train the proposed NRBM network model on five variants of Modified National Institute of Standards and Technology database (MNIST) benchmark dataset. The conducted experiments showed that the proposed NRBM is more robust to noisy data as compared to state-of-art approaches.**

Keywords*—*Multiclass classification; restricted Boltzmann machine; Polyak averaging; image classification; Modified National Institute of Standards and Technology Datasets

I. Introduction

Classification refers to the procedure of data taxonomy and data categorization in different forms, types and other distinctive classes. The aim of classification refers to the categorization, assembling, organization and differentiation of data into classes or groups. In term of statistics and machine learning, computer programs that are based on classification, supervisory learn from existing data that is input and categories it into different classes on behalf on learn data [1]. There are a few classification based problems, such as: bio metric classification, text classification, audio classification, and video classification [2], [3]. The phenomenon of generating results from classification or categorization is carried out either in two classes or more, which is called binary classification or multiclass classification respectively. Among classification algorithms, some are multiclass classifier by nature that allows the categorization of data in more than two classes. The algorithms with binary classification nature can also be used as multinomial classification using different approaches [4].

In multiclass classification, the aim of supervised behavior learning algorithms is to assign a desired class label for every input instance. For a given dataset of (a_j, b_j), where $b_j \in R_n$ is the j^th instance and $b_j \in 1, 2, 3...k$ is the j^th class label, the main purpose is to find a model for learning H, while $H(a_j) = b_j$ for all of the unknown testing instances. In the recent studies, many models and algorithms have been proposed for solving the two class classification problems, some of them are easy to extend for multiclass classification problems [5], [6], while some of them are based on special formulation in order to be able in solving multiclass classification problems. In these algorithms, there is a category of models in which the algorithms use different type of approaches for transforming multiclass classification into a bunch of binary classification problems in order to affectively solve it using binary classification approach [7], [8]. Another way of multiclass classification approaches is based on a try to pose a tree hierarchy on the output, additionally with the available class labels, in order to achieve a series of tests for new class labels detection.

Solving the multiclass classification problem is still exists as a challenging issue. In famous classifications approach, the multiclass classification problem is solved using the splitting of problem into many independent binary classification subproblems. Generally this type of solution is known as binarization, in which the classifier pretends to solve multiclass classification problem on the basis of binary classification [9]. The recent literature surveys shows that among all the multiclass classification problems, image classification attracted most of the current researchers interest.

Image classification based on unlabeled class oriented images data is a challenging problem in the field of computer vision. There is a need of classification model that is trained in an unsupervised behavior with less number of feature spaces. As there are various methods and techniques for image classification but still, it is an unneglectable challenge to design a classification model on unlabeled data. In this research, we are focusing to design and develop a normalized network based on restricted Boltzmann machine (RBM) [10] in addition with Polyak Averaging method [11] in training while updating parameters called NRBM. Since the learned features can be acquired economically at high scale yet comprises of temporal gestures because the images data contain temporally comprehensible frames.

In our conducted experiments, the proposed NRBM model has initiated the learning strategy from unlabeled data. We also showed that the proposed model can extract more useful features for final classification. Furthermore we conduct some experiments on MNIST variants benchmark datasets to validate our model. The results showed that the proposed NRBM model is able to learn the features that can be applied noisy and clear images. Moreover, a promising guide model for extracting and learning features to classify in a better way is designed. Managing and classification of huge amount of images by training the proposed NRBM model using unsupervised manner on the basis of low level features, is the main contribution of our research in this paper. The rest of paper is organized the following structure. Section 2 explores literature work on image classification models. Section 3 briefly describes the conventional restricted Boltzmann machine. Section 4 explains our proposed NRBM. In Section 5 we described our experiments to examine our proposed model and Section 6 presents the comparative analysis of NRBM with state-of-art models. Finally Section 7 concludes our research work and future directions. We believe our research will have a positive impact in many computer vision applications, e.g. Image recognition systems, video analyzers and image processing approaches.

II. Restricted Boltzmann Machine

Boltzmann Machine (BM) falls under the category of Artificial Neural Network (ANN) based on probability distribution for machine learning. Restricted Boltzmann Machine (RBM) is one of the famous variants of standard BM which was first created by Geoff Hinton [12]. The main purpose of RBM is reducing high dimensional data into low dimensional feature space. As it is a probability based approach that is why it is stochastic and generative in nature. The internal architecture of the RBM is similar to other Neural Networks (NNs) having layers with neurons in each layer. But in RBM there are only two layers. The first layer is referred to the input layer of the network, while the second layer is the hidden layer that is output from the input layer. There is a neural connection between the neurons of input layer and hidden layer. In standard BM, there exist connections between the neurons of the same layer but in RBM, there is a restriction that none of the neurons in same layer can communicate with the neurons between them.

Fig. 1 shows the internal architecture of RBM, in hidden layer there are x hidden nodes, while in visual layer has y number of visible nodes. The hidden nodes are represented by H_i and $i1, 2, 3, \ldots x$. While visible nodes are represented by V_j, such that $j1, 2, 3, \ldots y$. The weights connection between hidden and visible layers is represented by W_ij. The final equation for energy of RBM is:

$$\begin{aligned} Erbm(v, h|\phi) = -\sum_{i=1}^{x} v_i a_i - \sum_{j=1}^{y} hj \times \\ bj \sum_{i=1}^{x}\sum_{j=1}^{y} w_ij \times v_i \times h_j \end{aligned} \tag{1}$$

In Equation 1, $\phi = (W_ij, a_i, b_j)$, all these are real numbers. The bias value of hidden layer is denoted by b_j and the b_i represents the bais value of visible layer. The likelihood function that is also called the joint probability distribution of visible and hidden layer is formulated in Equation 2, if ϕ is known.

$$P(v, h|\phi) = \frac{1}{z} exp(-E(v, h|\phi)) \tag{2}$$

The normalization parameter $Z = (v, h)exp(-E(v, h|\phi))$. This visible and hidden layer's activation functions can be formatted in Equation 3 and Equation 4.

$$P(h_j, I|v) = sigmoid(\sum_{j}^{x} v_i \times w_ij + b_j) \tag{3}$$

$$P(v_i, I|v) = sigmoid(\sum_{i}^{y} h_j \times w_ij + a_i) \tag{4}$$

While $sigmoid(x) = \frac{1}{1+e_-x}$. Probability that the first j_th hidden node is activated is presented by Equation 3 while the activation of first i_th visible node is presented in Equation 4. Finding the best suitable parameter set that is possible for ϕ is the main purpose of RBM training. The input data can be bitterly fit by the trained model. For calculating the best value ofϕ using maximum likelihood function, we use the likelihood function based in log by giving training sample T_r, such that $T_r = v1, v2, v3, vTr$. The value of ϕ is:

$$\phi_* = argmaxL(\phi) = argmax \sum_{t_r=1}^{T_r} \log P(v^{tr}|\phi) \tag{5}$$

The step to step processing of RBM is given in the following Algorithm 1 and the conventional RBM architecture is shown in Fig. 1.

Psudocode Conventional RBM

```
Parameters initialization
    - Input feature set: $[x_1, x_2, x_3, \ldots, x_n]$
    - Weight Matrix: $W$
    - Bias vector of hidden layer:$b_j$
    - Bias vector of visible layer:$b_i$
    - The state of visible unit is set as training vector
Repeat
Visible units training
    - Compute $P(v, h|\phi)$ using Equation 2
    - Apply $P(v_i, I|v)$ of Equation 4
Hidden units training
    - Compute $P(v, h|\phi)$ using Equation 2
    - Apply $P(h_j, I|v)$ of Equation 3
    - Perform Gibbs sampling from hidden unit to visible unit using $h_1 i \in [0, 1]$
Update rule
    - Update weight $W$: $W = W + Wn$ while $Wn$ = Weight update factor
    - Update Bias $b_i$: $b_i = b_i + b_n$ while $b_n$ = Bias update factor for visible units
    - Update Bias $b_j$: $b_j = b_j + b_j n$ while $b_j n$ = Bias update factor for hidden units
    - Update energy function using Equation 1
While (All Layers Trained)
```

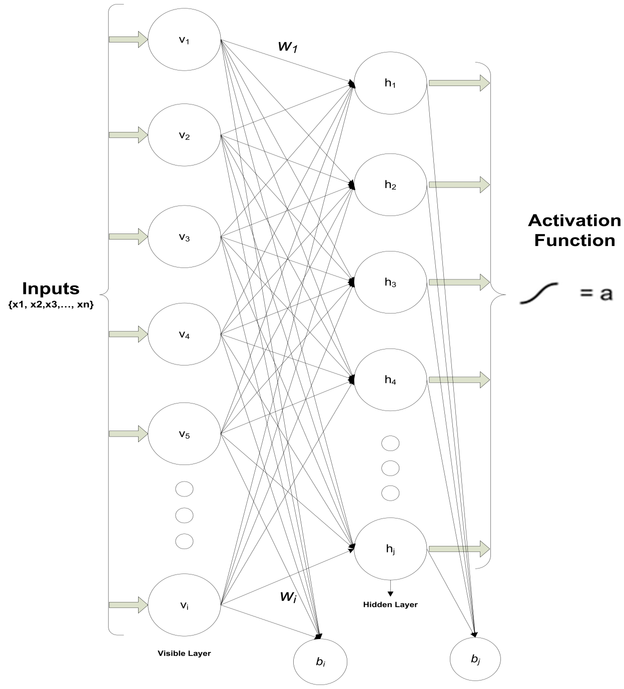


Fig. 1. Standard RBM internal architecture

III. RBM for Solving Multiclass Classification Problems

In the most recent research work RBM is widely used for different type of recognition, prediction and classification applications. For example [13] used RBM for face detection and recognition and get better results as compared to state-of-art methods. Particularly [14] conclude that confidently simple functions of stacked Restricted Boltzmann Machine (sRBM) can form a deep architecture for learning robust feature. Furthermore they trained convolutional Deep Boltzmann Machine (DBMs) [15] on faces image dataset (i.e., Caltech 101 images) and were impressive but unluckily their face detector can learn homogenous, aligned and a single category features.

In literature, BMs are been used more on prediction [16] and semantic based knowledge learning [17], [18] due the difficulty of generating output using low level representation, e.g. [19] split video frame sequence in chunks that describe actions in a discrete order. Author in [20] extract objects by removing background using the same manner in single domain boundary while [21] predict trajectories of human movement in a supervised way. Moreover there is also some found in order to model and forecast human to object [22] and human to human [23] connections based on supervised learning. In [24], authors have proposed Quantum Boltzmann Machine (QBM) inspired from the classical Boltzmann machines based one stochastic gradient descent in its training phase. They used a non-trivial training process for QBM and introduce bonds for the probabilities for quantum. The results have been verified by showing some examples that outperform conventional Boltzmann Machine based on training QBM with and without bounds using static diagonilazing.

In [25], the authors proposed a stochastic architecture based RBM digital images classification. They used four types of stochastic bit streams in their experiments, valued 512, 1024, 2048, 4096 of bit streams in order to achieve the balance between error rate and computation time. Stochastic number generator has been used for conversion of deterministic input data values to stochastic data values stream. In [26], the authors evaluate the ML models, specifically RBM for Anomaly Network Intrusion Detection Systems (A-NIDS). They applied hyperparameter tuning in RBM and report a few observations on the accuracy along with true positive and true negative rate. A balanced subset of ISCX benchmark dataset was used for training, validation and testing the proposed RBM model. Resultantly, RBM was found to be the better network for identification of attacks pattern in A-NIDS.

Gou C. presented hybrid discriminative restricted Boltzmann machines (HDRBMs) in [27] for car license plate recognition. He claim that HDRBM is the first hybrid model based on RBM that used for license number detection and recognition in such a wide view. They used two data subsets with high resolution images from car number plate dataset. The conducted experiment evaluated that their proposed HDRBM based on RBM outperform many of the state-of-art model for number plate detection and recognition. The claim that their model can be used efficiently in other countries as well but the layout for characters must predefine. In [28], the author presents another model known as discriminative Restricted Boltzmann Machine (discRBM) for learning discriminative feature set based on class labels. However their results conclude that discRBM performed well as compared to conventional RBM, deep RBM and SVM based on classification accuracy but still their model is limited to dataset size and complexity.

IV. Normalized Restricted Boltzmann Machine

Our proposed Normalized Restricted Boltzmann Machine NRBM is inspired by some state-of-art techniques and algorithm that are based on deep learning and unsupervised feature learning. The main source of inspiration is Restricted Boltzmann Machine (RBM) that is one of the famous variant of basic Boltzmann Machine (BM). The proposed model train on class label-less images for the purpose of extracting some amount of temporal informative features contain in the huge number test images data. To achieve this goal we studied some variants of RBM [29], [30], [31] and other classification algorithms [32], [33], [34]. After deep literature study we extended the restricted Boltzmann machine (RBM) from [10]. We present a two-phase training model that can easily extract background and foreground of the image separately. This foreground and background separation helps the model to learn the object feature inside the image easily and fast.

The main focus of this research model is tackling the multiclass classification problem that is caused by the high dimensional features space of large amount of data. We add Polyak averaging method with RBM in order to increase the speed of parameter convergence rate. We did not use traditional feature reduction techniques because the traditional techniques do not fulfill the user requirements for better feature extraction. In this paper, we proposed a deep learning approach based on normalized RBM called NRBM. The reason of using RBM is that, RBM uses high order structured statistical feature in training phase. Also the low dimension features is easily distinguishable from the features learnt by NRBM.

Taking the advantage of RBM network, the proposed model efficiently reduces the feature dimension space of images which is resultantly fall in increasing the accuracy of multiclass classification. In training phase of NRBM, we add Polyak Averaging to the first layer of RBM in order to get the normalized parameter set for next layer. We repeat the same procedure for each iteration in the proposed model until final classification. Equation 6, 7, 8 are the Polyak Averaging formulas for parameter updates.

$$w^n = w^n - 1 + \eta w^n \tag{6}$$

$$\bar{w}^n = \bar{w}^n - 1 - \frac{1}{n}(\bar{w}^n - 1 - w^n) \tag{7}$$

$$w^n = \bar{w}^n \tag{8}$$

Where, η is the scalar learning rate and $\bar{w}^n = \frac{1}{n}\sum_{i=1}^{x} w^i$.

Polyak Averaging increase the parameter convergence rate and also it does not disrupt the calculation training cost of the model because it comprises of only two simple addition functions. The sudden incline or decline in difference between two neighboring parameters for each layer is abolished by using averaging method. The set of feature obtained in result of the NRBM is given to final layer for final classification. Softmax classifier is considered as the final layer in the proposed NRBM model. The detailed step to step pseudocode of NRBM is given in the following Algorithm 2 and the network architecture is shown in Fig. 2. In the proposed model W is calculated using Equation 6 while $\bar{w}^n$ is solved using Equation 7.

Psudocode NRBM

```
Parameters initialization
    - Input feature set: $[x_1, x_2, x_3, \ldots, x_n]$
    - Weight Matrix: $W$
    - Bias vector of hidden layer:$b_j$
    - Bias vector of visible layer:$b_i$
    - The state of visible unit is set as training vector
Repeat
Visible units training
    - Compute $P(v, h|\phi)$ using Equation 2
    - Apply $P(v_i, I|v)$ of Equation 4
Hidden units training
    - Compute $P(v, h|\phi)$ using Equation 2
    - Apply $P(h_j, I|v)$ of Equation 3
    - Perform Gibbs sampling from hidden unit to visible unit using $h_1 i \in [0, 1]$
Update rule
    - Update weight using Equation 6 where $W$ is Weight update factor for visible units
    - Update Bias $b_i$: $b_i = b_i + b_n$ while $b_n$ = Bias update factor for visible units
    - Update weight using Equation 7 where $\bar{w}$ is Weight update factor for hidden units
    - Update Bias $b_j$: $b_j = b_j + b_j n$ while $b_j n$ = Bias update factor for hidden units
    - Update energy function using Equation 1
While (All Layers Trained)
```

V. Image Classification Steps using NRBM

Image classification is the famous application of classification in which images are being categorized on the basis of contents and object it consists [35]. Image classification is one of the famous problems in the field of classification. Like other classification approaches, the whole image classification phenomenon is also carried out in a few phases. Following are the common steps of image classification.

A. Image Pre-Processing

Image preprocessing refers to the procedures of processing both the input and output of the image on a very low level of abstraction [36]. The main purpose of image preprocessing is to enhance the image representation in order to cover the unwanted distortion and improves the quality of informative features for advance image processing. The whole image preprocessing is carried in a few steps, first the image data loading, secondly is the image resizing in order to fix all the image in a single standard size. The third step refers to the noise removal from the images and forth one is segmentation for the purpose of separating foreground and background objects.

VI. Features Extraction

In the feature extraction phase, the image is divided into segments and then each segment is processed in order to convert the pixels to their equal numerical values [37]. From these multi-dimensional values, important features are extracted from each segment one by one. Finally all the important features of the whole image which holds the complete representation of the entire image are extracted. The extracted features are the input of feature reduction module. Better feature extraction affects the feature reduction module positively.

A. Features Reduction

In the feature reduction phase of image processing, the features dimension of the image data is reduced from high dimension feature space. The main purpose of features reduction is to reduce the processing speed because the processing of high amount data for a classifier is time consuming task [38]. The proposed model mainly concern in effective feature reduction that ultimately results in the better image classification.

B. Classification

The last main stage of image classification is the use of a classifier for the final classification of image data. In the area machine learning, the problem of recognizing to which of a set of groups a new observation fits, based on a training dataset having observations or samples whose group association is known, referred as classification [39]. Different types of classifiers are used for classification. In the recent research studies, deep learning techniques are vastly used specifically for image classification.

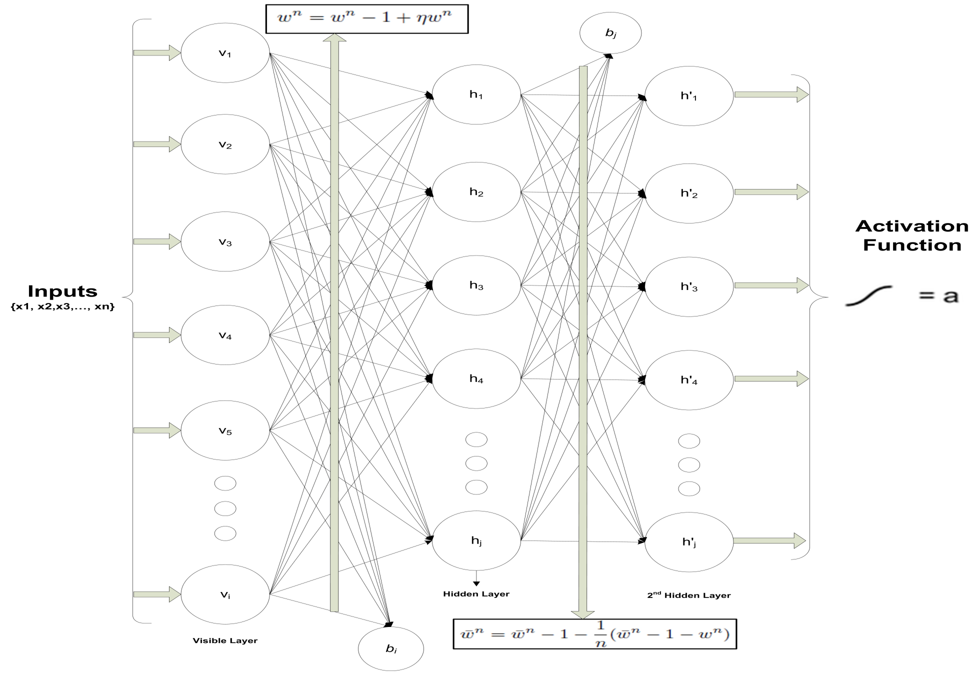


Fig. 2. NRBM internal architecture

The detailed factorial description of the proposed approach has been highlighted in Fig. 3.

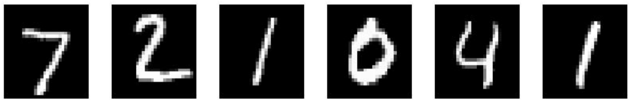


Fig. 4. MNIST small subset (mnist-basic)

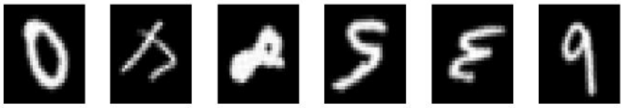


Fig. 5. MNIST random rotation digits (mnist-rot)

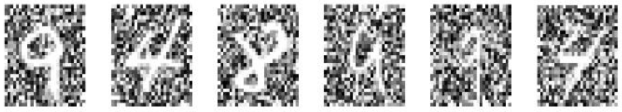


Fig. 6. MNIST random noise background digits (mnist-bg-rand)

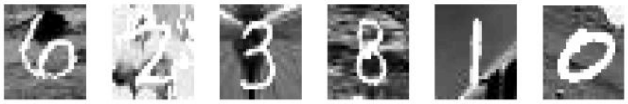


Fig. 7. MNIST random background digits (mnist-bg-img)

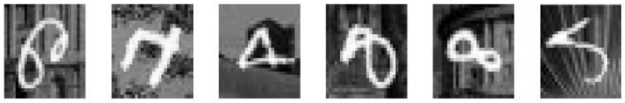


Fig. 8. MNIST rotation and image background digits (mnist-bg-img-rot)

VII. Experimental Results

We performed several experiments to verify the proposed NRBM model and evaluate the feature learning ability, which is the main improvement towards better classification. In this research, all the experiments were carried out on Intel core i5 cpu with 8GB of RAM having windows 10 operating system. The compiler and language used for developing and testing these algorithms is python3.6. For quick implementation of the proposed NRBM approach, an efficient numeric computational open source library Tensorflow [40] is used. It allows a simple and fast development for both CPU and GPU support. The dataset used for evaluating the proposed model are the

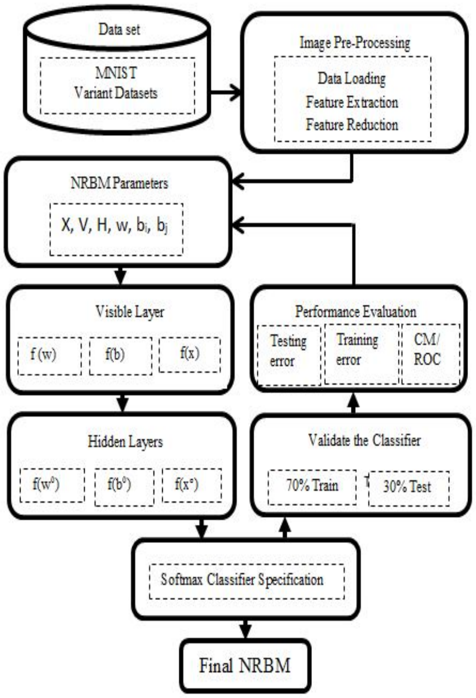


Fig. 3. Flow of image classification using NRBM

five variant subsets of benchmark data MNIST [41]. MNIST dataset contains 70,000 images of handwritten digits for 0 to 9. The size of each image is 28x28 pixels. In MINST the digits have been centered and the size of each digit is fixed to specific area inside the image. The variation subsets of MNIST are basic (small), random rotation digits (rot), random noise background digits (bg-rand), random background digits (bg-img) and rotation and image background digits (bg-img-rot). Fig. 4 to Fig. 8 are the randomly selected images from each of the variant subset, respectively. We select 5000 random images from each MNIST subset. The experiments were performed with two different testing and training ratios in order to verify the performance of NRBM. In first phase of experiments we select 50% of data for training and 50% for testing while in second phase we increase the training data ratio from 50% to 70% and reduce the testing data ratio from 50% to 30% of the whole dataset.

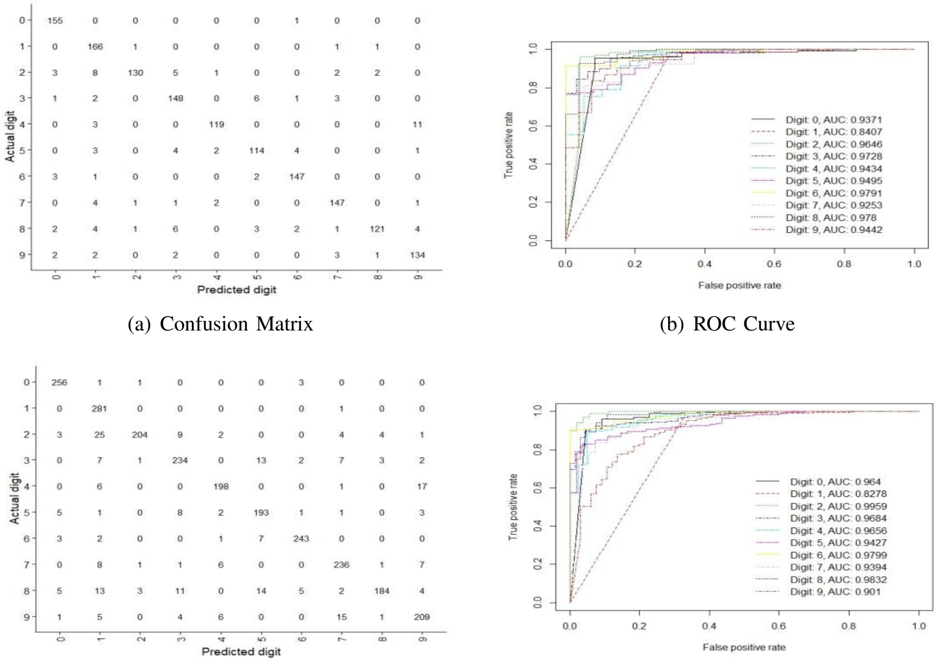
In these experiments, backpropagation approach is used in learning phase. There is no pre-processing or pre-training performed. For the proposed NRBM, the number of input neurons is 784. Learning rate is set to 0.03 for the hidden layer and we adjust the threshold for reconfiguration error to 0.3 after performing experiments with different thresholds rates.

After the backpropagation parameter adjustment we reduce the reconfiguration error threshold to 0.003. The number of output nodes from the final layer is 10 and the fine count for backpropagation is finally set to 500. In this research, the criteria for classification evaluation performance are Confusion Matrix (CM) and Receiver operating characteristic (ROC) curve. CM and ROC curve presents detail results of NRBM accuracy of correctly classified and misclassified instances at class level.

In order to evaluate the proposed NRBM model and its classification performance, Softmax classifier is used in the last layer of the network. Fig. 9 presents results of NRBM for MINST basic subset. Fig. 9(a) and 9(b) are representing CM and ROC curve for NRBM based on 70:30 training-testing dataset ratios while Fig. 9(c) and 9(d) for 50:50 of training-testing ratio. It is concluded easily that the results based on 70:30 ratio is are better than 50:50 training-testing ratio. Both of the CMs show that the classes with same pattern and similar features are mostly misclassified to each other e.g. class 4 and class 9, also 5 and 8. Fig. 10 represents the output of NRBM based on MINST rot variant subset. The accuracy per class is similar to MNIST basic subset, but the overall accuracy is decreased until 5% because the images are randomly rotated in this dataset which resultantly increased the complexity of the datasets.

Fig. 11 briefly describe the output of the proposed model for MINST bg-rand variant subset. The accuracy per class is better besides class 4 and class 2 that is misclassified to class 9 and class 7, respectively. This is because their features are quite similar. The ROC curve shows that there is a slowly paced decrease in the area under curve (AUC) for class 0 and 1 as compared to MNIST basic and rot.

Similarly, Fig. 12 shows the output of the NRBM model for MINST bg-img variant subset. The overall accuracy is improved for 70:30 training-testing ratio data, but at level, there is present some misclassification i.e. class 4 and class 9 that is misclassified to class 9 and class 7. The quite similar features are the reason of misclassification. Even with naked eye, handwritten 4 and 9 looks quite similar. The ROC shows that the AUC for class 1 and class 9 is lower as compared to other classes. Fig. 13 represents the results on MNIST bg-img-rot dataset. The accuracy per class is improved which is verified by the ROC in Fig. 13(b) and 13(d). From analysis point of view for all the experiment concluded from Fig. 9 to 13, there is gradual decrease can be seen in the overall classification accuracy of NRBM starting for MNIST basic subset until MNIST random background digits. This decrease is the reason of gradual increase in the complexity and noisiness in the datasets.



	0	1	2	3	4	5	6	7	8	9
0	155	0	0	0	0	0	1	0	0	0
1	0	166	1	0	0	0	0	1	1	0
2	3	8	130	5	1	0	0	2	2	0
3	1	2	0	148	0	6	1	3	0	0
4	0	3	0	0	119	0	0	0	0	11
5	0	3	0	4	2	114	4	0	0	1
6	3	1	0	0	0	2	147	0	0	0
7	0	4	1	1	2	0	0	147	0	1
8	2	4	1	6	0	3	2	1	121	4
9	2	2	0	2	0	0	0	3	1	134

(a) Confusion Matrix

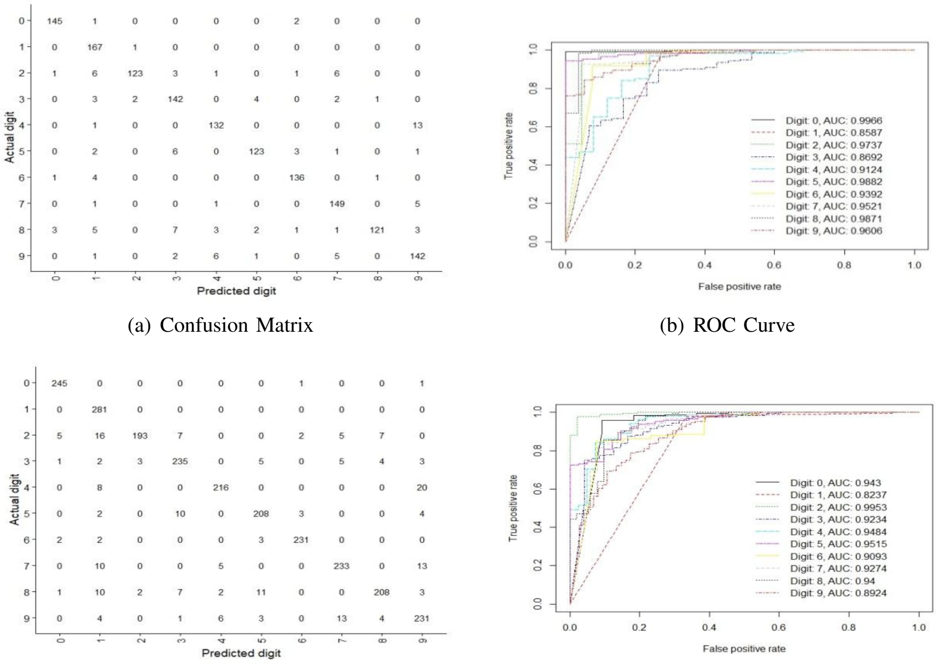
(b) ROC Curve

	0	1	2	3	4	5	6	7	8	9
0	256	1	1	0	0	0	3	0	0	0
1	0	281	0	0	0	0	0	1	0	0
2	3	25	204	9	2	0	0	4	4	1
3	0	7	1	234	0	13	2	7	3	2
4	0	6	0	0	198	0	0	1	0	17
5	5	1	0	8	2	193	1	1	0	3
6	3	2	0	0	1	7	243	0	0	0
7	0	8	1	1	6	0	0	236	1	7
8	5	13	3	11	0	14	5	2	184	4
9	1	5	0	4	6	0	0	15	1	209

(c) Confusion Matrix

(d) ROC Curve

Fig. 9. CM and ROC curve for 70:30 and 50:50 of training-testing ratio on MNIST small subset (mnist-basic)



	0	1	2	3	4	5	6	7	8	9
0	145	1	0	0	0	0	2	0	0	0
1	0	167	1	0	0	0	0	0	0	0
2	1	6	123	3	1	0	1	6	0	0
3	0	3	2	142	0	4	0	2	1	0
4	0	1	0	0	132	0	0	0	0	13
5	0	2	0	6	0	123	3	1	0	1
6	1	4	0	0	0	0	136	0	1	0
7	0	1	0	0	1	0	0	149	0	5
8	3	5	0	7	3	2	1	1	121	3
9	0	1	0	2	6	1	0	5	0	142

(a) Confusion Matrix

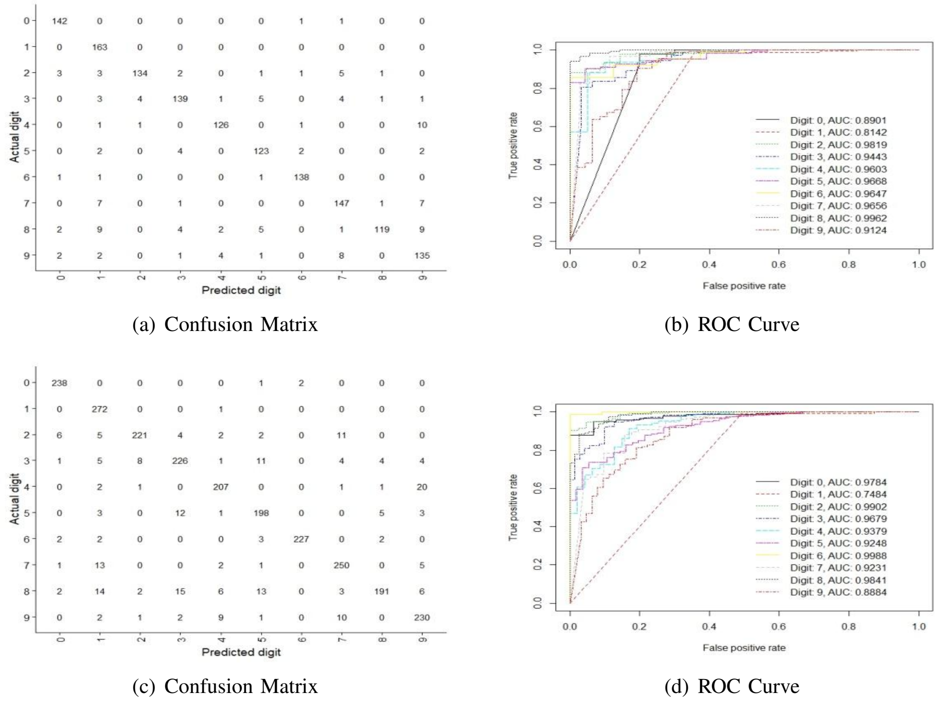
(b) ROC Curve

	0	1	2	3	4	5	6	7	8	9
0	245	0	0	0	0	0	1	0	0	1
1	0	281	0	0	0	0	0	0	0	0
2	5	16	193	7	0	0	2	5	7	0
3	1	2	3	235	0	5	0	5	4	3
4	0	8	0	0	216	0	0	0	0	20
5	0	2	0	10	0	208	3	0	0	4
6	2	2	0	0	0	3	231	0	0	0
7	0	10	0	0	5	0	0	233	0	13
8	1	10	2	7	2	11	0	0	208	3
9	0	4	0	1	6	3	0	13	4	231

(c) Confusion Matrix

(d) ROC Curve

Fig. 10. CM and ROC curve for 70:30 and 50:50 of training-testing ratio on MNIST random rotation digits (mnist-rot)



	0	1	2	3	4	5	6	7	8	9
0	142	0	0	0	0	0	1	1	0	0
1	0	163	0	0	0	0	0	0	0	0
2	3	3	134	2	0	1	1	5	1	0
3	0	3	4	139	1	5	0	4	1	1
4	0	1	1	0	126	0	1	0	0	10
5	0	2	0	4	0	123	2	0	0	2
6	1	1	0	0	0	1	138	0	0	0
7	0	7	0	1	0	0	0	147	1	7
8	2	9	0	4	2	5	0	1	119	9
9	2	2	0	1	4	1	0	8	0	135

(a) Confusion Matrix

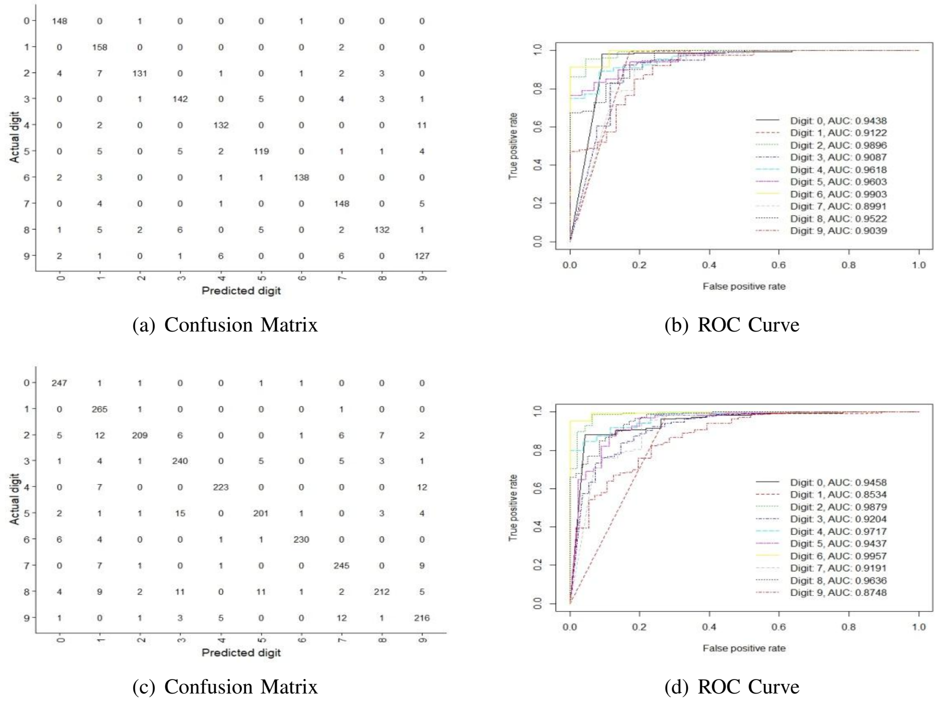
(b) ROC Curve

	0	1	2	3	4	5	6	7	8	9
0	238	0	0	0	0	1	2	0	0	0
1	0	272	0	0	1	0	0	0	0	0
2	6	5	221	4	2	2	0	11	0	0
3	1	5	8	226	1	11	0	4	4	4
4	0	2	1	0	207	0	0	1	1	20
5	0	3	0	12	1	198	0	0	5	3
6	2	2	0	0	0	3	227	0	2	0
7	1	13	0	0	2	1	0	250	0	5
8	2	14	2	15	6	13	0	3	191	6
9	0	2	1	2	9	1	0	10	0	230

(c) Confusion Matrix

(d) ROC Curve

Fig. 11. CM and ROC curve for 70:30 and 50:50 of training-testing ratio on MNIST random noise background digits (mnist-bg-rand)



	0	1	2	3	4	5	6	7	8	9
0	148	0	1	0	0	0	1	0	0	0
1	0	158	0	0	0	0	0	2	0	0
2	4	7	131	0	1	0	1	2	3	0
3	0	0	1	142	0	5	0	4	3	1
4	0	2	0	0	132	0	0	0	0	11
5	0	5	0	5	2	119	0	1	1	4
6	2	3	0	0	1	1	138	0	0	0
7	0	4	0	0	1	0	0	148	0	5
8	1	5	2	6	0	5	0	2	132	1
9	2	1	0	1	6	0	0	6	0	127

(a) Confusion Matrix

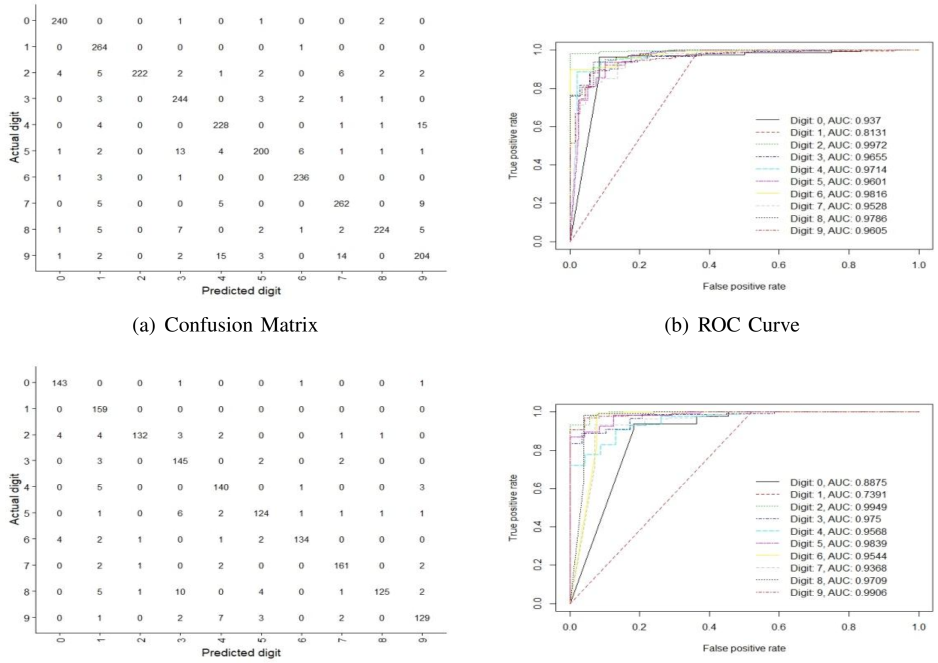
(b) ROC Curve

	0	1	2	3	4	5	6	7	8	9
0	247	1	1	0	0	1	1	0	0	0
1	0	265	1	0	0	0	0	1	0	0
2	5	12	209	6	0	0	1	6	7	2
3	1	4	1	240	0	5	0	5	3	1
4	0	7	0	0	223	0	0	0	0	12
5	2	1	1	15	0	201	1	0	3	4
6	6	4	0	0	1	1	230	0	0	0
7	0	7	1	0	1	0	0	245	0	9
8	4	9	2	11	0	11	1	2	212	5
9	1	0	1	3	5	0	0	12	1	216

(c) Confusion Matrix

(d) ROC Curve

Fig. 12. CM and ROC curve for 70:30 and 50:50 of training-testing ratio on MNIST random background digits (mnist-bg-img)



(a) Confusion Matrix

(b) ROC Curve

(c) Confusion Matrix

(d) ROC Curve

Fig. 13. CM and ROC curve for 70:30 and 50:50 of training-testing ratio on MNIST random background digits (mnist-bg-img-rot)

VIII. Comparative Analysis

A. Execution Time Comparison

The computational performance of all algorithms depends on the hardware, software and compilers. In our research, we used the same hardware and software combination that is mentioned in SECTION5. Results in Table I show the execution time comparison of the proposed NRBM model, Deep Belief Network (DBN) [42], Gated RBM (GRBM) [43], standard RBM [44] model. The overall performance of first two dataset between RBM GBM and NRBM is quite close. In MNIST basic, the execution time performance of RBM and GBM is better because the data is not complex relatively which causes feature extraction and feature reduction simpler and faster as compared to other datasets. As the data is more and noisy and complex in MNIST sub-bg-rand, sub-bg-img and sub-bg-img-rot, the execution time performance of NRBM is better as compared to DBN, GBM, and standard RBM.

B. Training Error Rate Comparison

The most crucial part in any AI model is in the training phase. In the conducted experiments, the same parameters tuning is been followed as mentioned in Section 5. 5000 random images from each dataset are selected for training the comparative models. The back propagation behavior is applied in training phase. Table II presents the training error rate comparative analysis for NRBM with some state-of-art models. In the concluded training error rates, firstly the DBN performed well for MNIST basic dataset. In sub-rot dataset, RBM is better for 70/30 ratio while NRBM proved to be better for 50/50 ratio of training-testing data. In the rest datasets experiments, RBM and NRBM are performed relatively close to each other but in the overall results NRBM outperformed the comparative models.

C. Testing Error Rate Comparison

The experiments conducting for test error rate analysis are given in Table III. The classification accuracy and performance depends on the internal architecture of the model and the classifier used for final classification. In order to conduct fair experiments, we used the same parameter settings and followed architecture for all the models with Softmax layer for the final classification. Like training comparative analysis, we also test all the models with a two training-testing data ratio. The results in Table III conclude that the testing error of DBN is lower for MNIST basic subset and RBM for sub-rot subset because the dataset complexity is very low in these two subsets. Though the error rate of RBM is close to NRBM but still in sub-bg-rand, sub-bg-img and bg-img-rot dataset, the proposed NRBM outperformed DBN, GBM and RBM.

IX. Conclusion

RBM is one of the most usable models in the area of deep learning. This paper proposed a NRBM that is based on Polyak Averaging method in training, in order to improve the weights and bais update parameters function. This research is carried out with three main contributions. Firstly using NRBM to learn features and reduce feature dimension efficiently. Secondly we use Softmax classifier in the last layer of the proposed model for final classification. Lastly we conduct sufficient experiments for comparative analysis to verify the significance and effectiveness of the proposed NRBM. The normalization function in RBM makes the model faster in learning since Polyak averaging method is composed of simple addition

TABLE I. Execution time comparison of different models on MNIST variant-datasets

Datasets	Sub-Basic		Sub-Rot		Sub-Bg-rand		Sub-Bg-img		Sub-Bg-img-rot	
Training/Testing	50/50	70/30	50/50	70/30	50/50	70/30	50/50	70/30	50/50	70/30
DBN	15m 55s	18m 23s	20m 26s	23m 42s	22m 56s	27m 02s	26m 36s	32m 43s	33m 55s	41m 24s
GBM	12m 32s	16m 46s	20m 50s	22m 28s	20m 10s	24m 05s	-	-	-	-
RBM	12m 15s	14m 52s	17m 44s	21m 36s	18m 26s	22m 12s	24m 57s	29m 03s	29m 45s	34m 28s
NRBM	14m 54s	15m 35s	17m 35s	19m 25s	19m 02s	21m 36s	21m 10s	23m 48s	25m 28s	28m 33s

TABLE II. Training error rate comparison of different models on MNIST variant-datasets

Datasets	Sub-Basic		Sub-Rot		Sub-Bg-rand		Sub-Bg-img		Sub-Bg-img-rot	
Training/Testing	50/50	70/30	50/50	70/30	50/50	70/30	50/50	70/30	50/50	70/30
DBN	2.96	3.63	7.68	10.25	11.53	15.85	17.52	21.12	23.40	28.65
GBM	8.78	9.31	13.85	19.12	19.95	23.45	-	-	-	-
RBM	6.64	7.20	7.94	8.98	10.05	10.82	14.94	17.87	28.32	36.52
NRBM	3.10	3.78	6.05	9.06	9.20	10.10	15.10	17.02	21.60	24.85

TABLE III. Testing error rate comparison of different models on MNIST variant-datasets

Datasets	Sub-Basic		Sub-Rot		Sub-Bg-rand		Sub-Bg-img		Sub-Bg-img-rot	
Training/Testing	50/50	70/30	50/50	70/30	50/50	70/30	50/50	70/30	50/50	70/30
DBN	9.58	7.43	14.96	12.10	18.14	22.74	21.87	23.70	27.25	33.85
GBM	15.54	13.31	18.10	19.12	24.68	26.34	-	-	-	-
RBM	14.78	10.08	14.17	11.19	15.85	17.70	18.63	22.12	42.32	38.12
NRBM	12.96	9.45	16.32	13.40	15.21	16.50	18.92	20.78	24.48	29.86

function. Python based efficient computational library named Tensorflow, is used for implementation. The experiments are carried out on five MNIST variant datasets. As a conclusion the overall result revealed that the proposed model performed better when the complexity of dataset increased. Our future goals is to move forward our research by incorporating more optimization and normalization methods to the standard RBM and also applying the proposed model to more complex and big datasets.

Acknowledgment

The authors would like to thank Ministry of Higher Education (MOHE) Malaysia and Universiti Tun Hussein Onn Malaysia (UTHM) for financially supporting this Research under FRGS grants vote no. 1641.

References

[1] R. Kune, P. K. Konugurthi, A. Agarwal, R. R. Chillarige, and R. Buyya, “The anatomy of big data computing,” *Software: Practice and Experience*, vol. 46, no. 1, pp. 79–105, 2016.

[2] Y. Sun, M. Zhang, Z. Sun, and T. Tan, “Demographic analysis from biometric data: Achievements, challenges, and new frontiers,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 2, pp. 332–351, 2018.

[3] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very deep convolutional networks for text classification,” *arXiv preprint arXiv:1606.01781*, 2016.

[4] M. Liu, D. Zhang, S. Chen, and H. Xue, “Joint binary classifier learning for ecoc-based multi-class classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2335–2341, 2016.

[5] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural Networks*, vol. 106, pp. 249–259, 2018.

[6] J. A. Nasiri, N. M. Charkari, and S. Jalili, “Least squares twin multi-class classification support vector machine,” *Pattern Recognition*, vol. 48, no. 3, pp. 984–992, 2015.

[7] P. Xu, F. Davoine, H. Zha, and T. Denoeux, “Evidential calibration of binary svm classifiers,” *International Journal of Approximate Reasoning*, vol. 72, pp. 55–70, 2016.

[8] E. Vissol-Gaudin, A. Kotsialos, C. Groves, C. Pearson, D. A. Zeze, and M. C. Petty, “Computing based on material training: Application to binary classification problems,” in *2017 IEEE International Conference on Rebooting Computing (ICRC)*. IEEE, 2017, pp. 1–8.

[9] A. Sen, M. M. Islam, K. Murase, and X. Yao, “Binarization with boosting and oversampling for multiclass classification,” *IEEE transactions on cybernetics*, vol. 46, no. 5, pp. 1078–1091, 2016.

[10] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 791–798.

[11] W. Sun, D. H. Poot, X. Yang, W. J. Niessen, and S. Klein, "Averaged stochastic optimization for medical image registration based on variance reduction," in *International Workshop on Biomedical Image Registration*. Springer, 2018, pp. 69–79.

[12] G. E. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 599–619.

[13] T. D. Nguyen, T. Tran, D. Phung, and S. Venkatesh, "Tensor-variate restricted boltzmann machines," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[14] J. Chu, H. Wang, H. Meng, P. Jin, and T. Li, "Restricted boltzmann machines with gaussian visible units guided by pairwise constraints," *IEEE transactions on cybernetics*, no. 99, pp. 1–14, 2018.

[15] Z. Gan, R. Henao, D. Carlson, and L. Carin, "Learning deep sigmoid belief networks with data augmentation," in *Artificial Intelligence and Statistics*, 2015, pp. 268–276.

[16] R. Hrasko, A. G. Pacheco, and R. A. Krohling, "Time series prediction using restricted boltzmann machines and backpropagation," *Procedia Computer Science*, vol. 55, pp. 990–999, 2015.

[17] S. Tsogkas, I. Kokkinos, G. Papandreou, and A. Vedaldi, "Semantic part segmentation with deep learning," *arXiv preprint arXiv:1505.02438*, vol. 3, no. 7, 2015.

[18] N. Kaur, G. Kunapuli, T. Khot, K. Kersting, W. Cohen, and S. Natarajan, "Relational restricted boltzmann machines: A probabilistic logic learning approach," in *International Conference on Inductive Logic Programming*. Springer, 2017, pp. 94–111.

[19] H. Vu, T. D. Nguyen, and D. Phung, "Detection of unknown anomalies in streaming videos with generative energy-based boltzmann models," *arXiv preprint arXiv:1805.01090*, 2018.

[20] S. Lee and D. Kim, "Background subtraction using the factored 3-way restricted boltzmann machines," *arXiv preprint arXiv:1802.01522*, 2018.

[21] M. Raj, V. B. Semwal, and G. Nandi, "Bidirectional association of joint angle trajectories for humanoid locomotion: the restricted boltzmann machine approach," *Neural Computing and Applications*, vol. 30, no. 6, pp. 1747–1755, 2018.

[22] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4346–4354.

[23] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom, "Deep representation learning for human motion prediction and classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6158–6166.

[24] M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko, "Quantum boltzmann machine," *Physical Review X*, vol. 8, no. 2, p. 021050, 2018.

[25] B. Li, M. H. Najafi, and D. J. Lilja, "Using stochastic computing to reduce the hardware requirements for a restricted boltzmann machine classifier," in *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 2016, pp. 36–41.

[26] T. Aldwairi, D. Perera, and M. A. Novotny, "An evaluation of the performance of restricted boltzmann machines as a model for anomaly network intrusion detection," *Computer Networks*, vol. 144, pp. 111–119, 2018.

[27] C. Gou, K. Wang, Y. Yao, and Z. Li, "Vehicle license plate recognition based on extremal regions and restricted boltzmann machines," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1096–1107, 2016.

[28] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 536–543.

[29] S. Nie, Z. Wang, and Q. Ji, "A generative restricted boltzmann machine based method for high-dimensional motion data modeling," *Computer Vision and Image Understanding*, vol. 136, pp. 14–22, 2015.

[30] K. Cho, A. Ilin, and T. Raiko, "Improved learning of gaussian-bernoulli restricted boltzmann machines," in *International conference on artificial neural networks*. Springer, 2011, pp. 10–17.

[31] V. Mnih, H. Larochelle, and G. E. Hinton, "Conditional restricted boltzmann machines for structured output prediction," *arXiv preprint arXiv:1202.3748*, 2012.

[32] Z. Zhang, Zhang, and Khelifi, *Multivariate Time Series Analysis in Climate and Environmental Research*. Springer, 2018.

[33] J. Tanha, M. van Someren, and H. Afsarmanesh, "Semi-supervised self-training for decision tree classifiers," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 1, pp. 355–370, 2017.

[34] L. Gao, J. Li, M. Khodadadzadeh, A. Plaza, B. Zhang, Z. He, and H. Yan, "Subspace-based support vector machines for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 2, pp. 349–353, 2015.

[35] P. Druzhkov and V. Kustikova, "A survey of deep learning methods and software tools for image classification and object detection," *Pattern Recognition and Image Analysis*, vol. 26, no. 1, pp. 9–15, 2016.

[36] F. Albu and A. Zamfir, "Image processing method and apparatus," Mar. 24 2015, uS Patent 8,989,516.

[37] Z. Fan, D. Bi, L. He, M. Shiping, S. Gao, and C. Li, "Low-level structure feature extraction for image processing via stacked sparse denoising autoencoder," *Neurocomputing*, vol. 243, pp. 12–20, 2017.

[38] S. De, M. Singha, K. Kumari, R. Selot, and A. Gupta, "Dimension reduction using image transform for content-based feature extraction," in *Feature Dimension Reduction for Content-Based Image Identification*. IGI Global, 2018, pp. 26–40.

[39] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[40] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[41] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 473–480.

[42] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[43] X. Jin, T. He, C. Wan, L. Yi, G. Ding, and D. Shen, "Automatic gating of attributes in deep structure." in *IJCAI*, 2018, pp. 2305–2311.

[44] M.-A. Côté and H. Larochelle, "An infinite restricted boltzmann machine," *Neural computation*, vol. 28, no. 7, pp. 1265–1288, 2016.

Key Schedule Algorithm using 3-Dimensional Hybrid Cubes for Block Cipher

Muhammad Faheem Mushtaq[1], Sapiee Jamel[2], Siti Radhiah B. Megat[3], Urooj Akram[4], Mustafa Mat Deris[5]
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, 86400 Batu Pahat, Johor, Malaysia[1, 2, 3, 5]
Faculty of Computer Science and Information Technology
Khwaja Fareed University of Engineering and Information Technology, 64200 Rahim Yar Khan, Pakistan[1, 4]

***Abstract*—A key scheduling algorithm is the mechanism that generates and schedules all session-keys for the encryption and decryption process. The key space of conventional key schedule algorithm using the 2D hybrid cubes is not enough to resist attacks and could easily be exploited. In this regard, this research proposed a new Key Schedule Algorithm based on coordinate geometry of a Hybrid Cube (KSAHC) using Triangular Coordinate Extraction (TCE) technique and 3-Dimensional (3D) rotation of Hybrid Cube surface (HCs) for the block cipher to achieve large key space that are more applicable to resist any attack on the secret key. The strength of the keys and ciphertext are tested using the Hybrid Cube Encryption Algorithm (HiSea) based on Brute Force, entropy, correlation assessment, avalanche effect and NIST randomness test suit which proves the proposed algorithm is suitable for the block cipher. The results show that the proposed KSAHC algorithm has performed better than existing algorithms and we remark that our proposed model may find potential applications in information security systems.**

***Keywords*—Encryption; decryption; key schedule algorithm; hybrid cube; block cipher**

I. Introduction

Security is the major concern due to the fast growth of the internet in today's digital world and it is important to provide security of data from unauthorized access [1]. Hence, secure communication is the basic requirement of every transaction over networks. Cryptography is an important component to ensure secure communication of data by using the security services like confidentiality, authentication, data integrity and non-repudiation. Data confidentiality refers to the protection of sensitive data from being accessed by unauthorized parties [2]. Traditionally, the cryptographic algorithms comprise of different mathematical and logical components integrated together as part of the algorithms [3], [4], [5]. The development of the fully secured cryptographic algorithm is difficult due to the challenges from cryptanalysts who continuously trying to break any available cryptographic systems. So, the selection of the right cryptographic algorithm is essential to accomplish the high-security requirements to ensure the protection of cryptographic components from cryptanalysis [6], [7]. In this regard, a key schedule algorithm is employed to generate secret keys and plays an important role in the development of encryption schemes. In order to resist the related key attack, many types of research were conducted to develop a powerful and significant key generation algorithm that increases the difficulty for a cryptanalyst to recover the secret key [8]. All cryptographic algorithms are recommended to follow 128, 192 and 256-bits key lengths proposed by Advanced Encryption Standard (AES) [9].

Permutation plays an important role in the development of cryptographic algorithms and it contains the finite set of numbers or symbols that are used to mix up a readable message into ciphertext as shown in transposition cipher [10]. The logic behind any cryptographic algorithm is the number of possible combinations in the key space, and bigger key-space could be achieved from the used of flips and twists of the elements of the cube which ensure every state of the cube is actually permuted [11], [12]. An image permutation algorithm based on the geometrical projection and shuffling in the design of key schedule algorithm is used to increases the security of original image by preventing it from outside attacks [13], [14]. The computing information about complex geometric primitives is often costly, while computational geometry determines many asymptotically effective algorithms for such problems are indicated [15]. It mentioned that if the coordinate system is adopted in the plane then key quantities can be circularly permuted. The rotation of the object is used to shift the position along a circular path in the *xy* plane. Whereas, the translation of an object is employed to shift the position along the straight path from one coordinate to another [16], [17]. The construction of magic cubes using the concept of a magic square and two orthogonal Latin squares was proposed [18]. The magic cube is 3-Dimensional (3D) coordinates consisting of six faces that are used in the development of complex permutation as apart of the design of cipher. The cubes rotation technique is applied to the image pixels to produce an encrypted picture [19] and revert the rotation to decrypt the image. Moreover, the image scrambling technique using the rotation of rows and columns of the magic cube is used to break the relationship between the image elements which creates the encryption [20]. Similarly, the cube rotation technique has been applied to encrypt text which provides more security and efficiency [21] compared to other operations.

Hybrid cubes are generated on the basis of Latin squares, Orthogonal Latin Squares, Magic Squares and Magic Cubes [7]. Based on that, a new way was found for further development of new transformation based on a permutation of integer numbers and develops a non-binary block cipher. Furthermore, a non-binary block cipher is proposed using all possible combination of 2-Dimensional (2D) hybrid cubes as the source for the encryption and decryption keys [22]. The Rajavel's encryption algorithm and the cubical key are generated using the hybridization and rotation of hybrid cubes by shuffling the cubes [23]. Similarly, the HiSea encryption algorithm performed hybridization with 2D hybrid cubes which are a very time-consuming process and it generates the key space that is not sufficient to resist attacks and could easily be exploited. This research opens a new way for creating a key schedule algorithm using 3D hybrid cubes based on permutation and combination of integer numbers.

In this regard, this research proposed a new Key Schedule Algorithm based on the coordinate geometry of a Hybrid Cube (KSAHC) for HiSea encryption algorithm [22]. The KSAHC transformation using Three-Dimensional (3D) hybrid cube is to create a large key space that will be used as encryption and decryption keys which makes the Brute Force attack difficult and time-consuming. KSAHC encryption keys are represented as an $n \times n \times n$ matrix of integer numbers and used in the development of the permutation and substitution of order 4 square matrix. A new cube structure based on the Triangular Coordinate Extraction [24] and rotation of Hybrid Cube surface [25] is proposed where the layer entries are between a set of integers from 1618 to 11198. TCE technique is used to extract the coordinate of hybrid cube during the rotation and it plays an important role in the development of KSAHC algorithm. Also, this transformation will be able to generate random (4×4×4) matrices from any hybrid cube layers. Furthermore, the proposed algorithm has been implemented into existing non-binary block cipher and observes the promising impact of keys on the ciphertext. Hence, by using the proposed KSAHC algorithm for HiSea, this research leads to the enhancement of the strength and validation of encryption key and the cipher. The contributions and novelty of this study are as follows:

- To propose a new Key Schedule Algorithm based on coordinate geometry of a Hybrid Cubes (KSAHC) for the non-binary block cipher.
- To overcome the problem of small key space that could occur due to the 2D hybrid cube layers, a new approach has been adopted based on the 3D rotation of HCs by using the columns and rows shift transformation.
- A comparative analysis of the proposed algorithm with AES, HiSea, and DKSA has been completed, and the strength of the encryption keys and ciphertext are examined based on Brute Force, entropy, correlation assessment, avalanche effect, and NIST randomness test suit.
- The novelty of this research is to incorporate the concept of coordinate geometry, 3D rotation of *HCs* and TCE technique into the key scheduling algorithm.

The remaining paper is organized as follows: Section II discusses the material and methods in which explain the detail design of the proposed Key Schedule Algorithm based on coordinate geometry of Hybrid Cube (KSAHC) for the non-binary block cipher. Section III explains the results and discussion of the proposed algorithm. Section IV presents the conclusion and future work of this research.

II. Material and Methods

This section discusses the design of the new Key Schedule Algorithm based on coordinate geometry of a Hybrid Cubes (KSAHC) for HiSea encryption algorithm [22].

The schematics of a conventional HiSea Encryption algorithm uses individual cubes with a 2D structure for the development of key schedule algorithm but the proposed KSAHC algorithm has been developed using the 3D structure of hybrid cube and its rotation using ShiftColumn and ShiftRow transformation as shown in Fig. 1. The first step in the development of KSAHC is to generate rotation of *HCs* which is the main element of the construction of encryption and decryption key in the key schedule algorithm. The shuffling and mixing of rows and columns of hybrid cube provide the resultant matrices that must be invertible. These invertible matrices are used in the development of encryption and decryption keys in the symmetric block cipher. The second step develops a key scheduling algorithm which involves the generation of the key table using the matrices that are utilized in the rotation of the hybrid cube. Previously, the inner matrix multiplication of magic cubes is used to generates hybrid cubes and these cubes are used to generate a key table in the key schedule algorithm. The third step involves encryption of messages using the encryption key to form ciphertext that comprises of integer numbers string employed in the construction of non-binary block cipher and final step is to decrypt the message to form an encryption algorithm. For this purpose, the detailed design of the key schedule algorithm using the triangular 3D rotation of *HCs*, HiSea encryption, and decryption algorithms are discussed. Based on Fig. 1, the entire elements in the dotted box illustrate the components of KSAHC framework. The rotation is needed in transforming the ciphertext to plaintext to obtain the original plaintext from the ciphertext.

A. Proposed KSAHC Algorithm

The proposed key schedule algorithm undergoes through the following main phases; generation of key matrices using TCE technique, ShiftColumns, ShiftRows, unique matrices operation, and triangular key matrices. Using the layers of hybrid cubes [22], a new cube structure is generated that are used to computationally secures the encryption and decryption process of existing HiSea encryption algorithm. Every algorithm is almost having inputs, processes, and output. The input to the proposed key schedule algorithm is the user password. The KSAHC algorithm determines a permutation based on user password in generating the triangular key matrices from all faces of the hybrid cube. The Initial Vector (IV) is also used as the secondary security measure with the user password to prevent repetition in keys. The process of 3D rotation of *HCs* can be explained in Fig. 2.

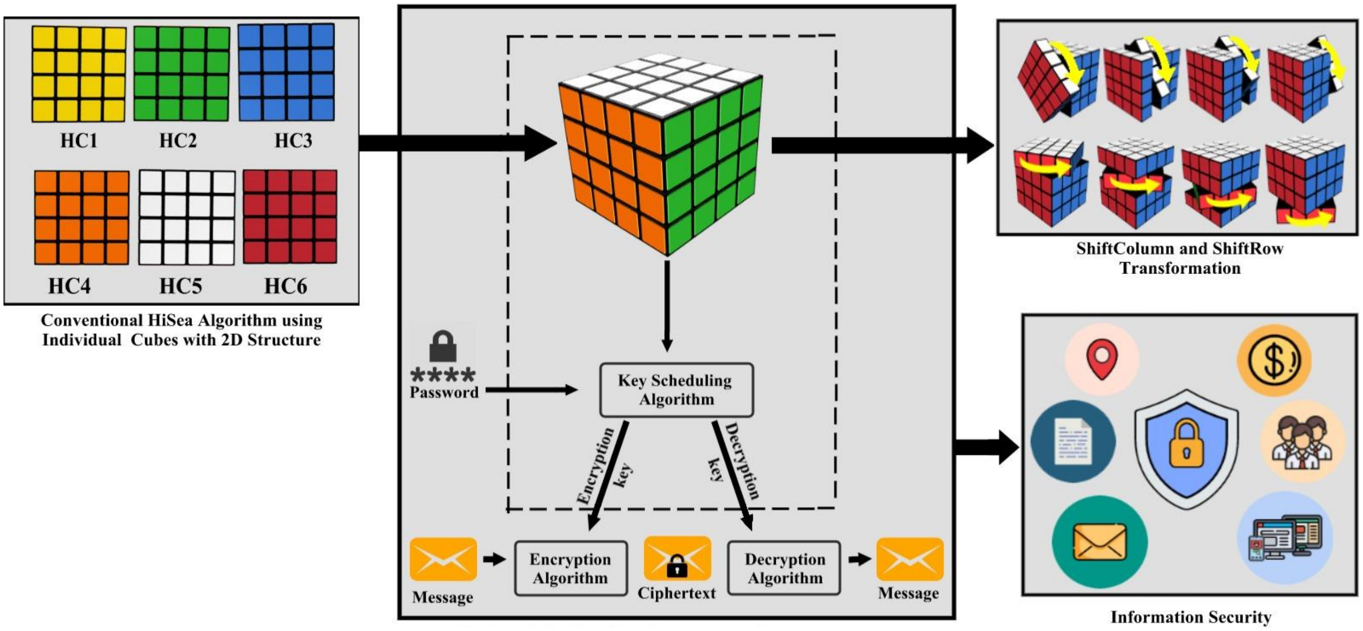


Fig. 1. Framework of the Proposed KSAHC Algorithm.

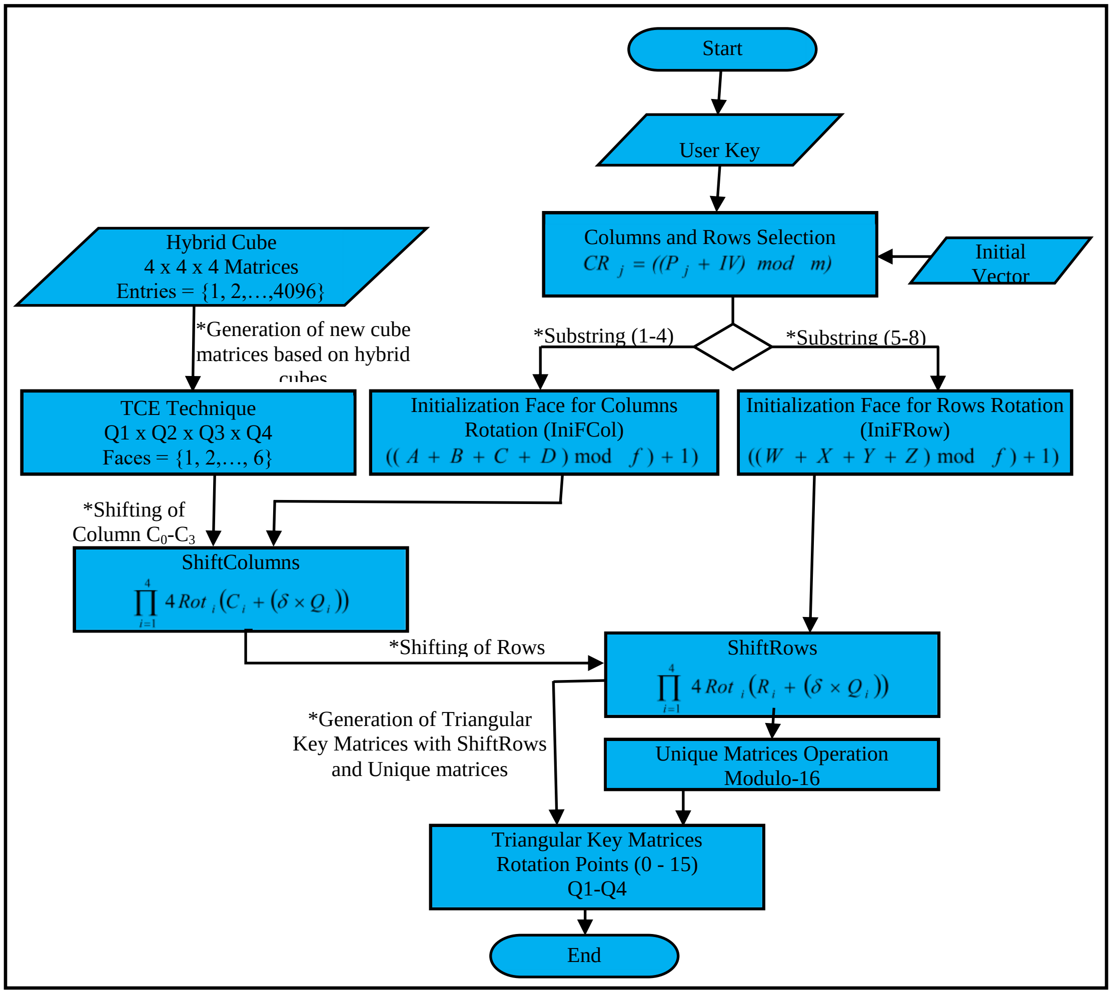

Start
User Key
Columns and Rows Selection
$CR_j = ((P_j + IV) \bmod m)$
Initial Vector
Hybrid Cube
4 x 4 x 4 Matrices
Entries = {1, 2,…,4096}
*Generation of new cube matrices based on hybrid cubes
*Substring (1-4)
*Substring (5-8)
TCE Technique
Q1 x Q2 x Q3 x Q4
Faces = {1, 2,…, 6}
Initialization Face for Columns Rotation (IniFCol)
$((A + B + C + D) \bmod f) + 1)$
Initialization Face for Rows Rotation (IniFRow)
$((W + X + Y + Z) \bmod f) + 1)$
*Shifting of Column C_0-C_3
ShiftColumns
$\prod_{i=1}^{4} 4\,Rot_i(C_i + (\delta \times Q_i))$
*Shifting of Rows
ShiftRows
$\prod_{i=1}^{4} 4\,Rot_i(R_i + (\delta \times Q_i))$
*Generation of Triangular Key Matrices with ShiftRows and Unique matrices
Unique Matrices Operation Modulo-16
Triangular Key Matrices Rotation Points (0 - 15) Q1-Q4
End

Fig. 2. The Process of KSAHC Algorithm.

The 8-digit value of the selection of Columns and Rows (CR_j) is employed for the rotation of columns and rows of the hybrid cubes. Furthermore, the CR_j is also used to identify the initialization faces of columns and rows. ShiftColumns and ShiftRows operations are performed on different faces of the hybrid cube and TCE technique is demonstrated. The 3D rotation pattern of both operations is depending on the CR_j value. A new combination of triangular key matrices is generated by using the ShiftRows operation and the rotation points based on the unique matrices. This new combination of layer entries using the rotation of *HCs* can be used to add randomness in the output of the proposed algorithm. Only the invertible triangular key matrices of the hybrid cube are used as the encryption and decryption keys in the non-binary block cipher. Furthermore, the encryption and decryption algorithm of HiSea encryption algorithm is also discussed in order to test the strength of the proposed algorithm. The detailed design of these components will be explained in the following sub-section.

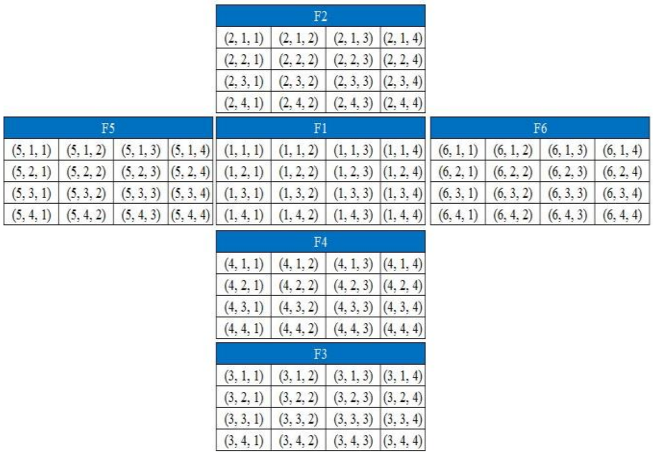
1) Generation of key matrices: The input key matrices (also referred to as the states) are structured into the 4 × 4 matrices. In this regard, 24 key matrices are generated using the HiSea encryption algorithm [22] and the TCE technique is demonstrated. By using the TCE technique, each key matrix from HiSea key table is used to generate one row in the new cube matrix. Similarly, four key matrices are used to generate one face of the hybrid cube. Finally, 24 key matrices are employed to generate six key matrices that are used in the rotation of HCs as shown in Fig. 3.

a) Design of Triangular Coordinate Extraction (TCE) Technique: The design of HCs is divided into six faces and each face is further divided into four quarters (Q1, Q2, Q3, Q4) by intersection of two diagonal lines pass through the center of a circle [24]. The primary diagonal lies on the x-axis while the secondary diagonal is on the y-axis as presented in Fig. 4. The element of the coordinates of the hybrid cube shown as f, i and j in which the element f shows the six faces of the hybrid cube, while i represents the rows and the j represents the columns of the hybrid cube. Firstly, the rotation of HCs of face 1 is considered and after that, a similar process is applied in the other faces. The rotation of triangular HCs is counterclockwise which is the main component in the generation of the encryption keys. The rotation points 0 to 15 around the quarters Q1 to Q4 represent the position of coordinates in which the Q1 is the passage from rotation points 0 to 4, Q2 lies between 4 to 8, Q3 lies from 8 to 12 and Q4 is the passage from 12 to 0.

Definition 1. Let the *HCs* be 4 × 4 × 4 square matrices, then the center of *HCs* is calculated by using the intersection of primary and secondary diagonals and it is possible if the elements of coordinates of rows and columns satisfy the reflexive and symmetric properties [25]. The properties of diagonals for the six faces of the hybrid cube are as follows:

2) Primary diagonal is defined as the collection of entries *HCs* (*f*, *i*, *j*), where $i = j$. The coordinates of primary diagonal for each face (*f*) of hybrid cube comprises the follows:

{(1, 1), (2, 2), (3, 3), (4, 4)}



F2			
(2, 1, 1)	(2, 1, 2)	(2, 1, 3)	(2, 1, 4)
(2, 2, 1)	(2, 2, 2)	(2, 2, 3)	(2, 2, 4)
(2, 3, 1)	(2, 3, 2)	(2, 3, 3)	(2, 3, 4)
(2, 4, 1)	(2, 4, 2)	(2, 4, 3)	(2, 4, 4)

F5			
(5, 1, 1)	(5, 1, 2)	(5, 1, 3)	(5, 1, 4)
(5, 2, 1)	(5, 2, 2)	(5, 2, 3)	(5, 2, 4)
(5, 3, 1)	(5, 3, 2)	(5, 3, 3)	(5, 3, 4)
(5, 4, 1)	(5, 4, 2)	(5, 4, 3)	(5, 4, 4)

F1			
(1, 1, 1)	(1, 1, 2)	(1, 1, 3)	(1, 1, 4)
(1, 2, 1)	(1, 2, 2)	(1, 2, 3)	(1, 2, 4)
(1, 3, 1)	(1, 3, 2)	(1, 3, 3)	(1, 3, 4)
(1, 4, 1)	(1, 4, 2)	(1, 4, 3)	(1, 4, 4)

F6			
(6, 1, 1)	(6, 1, 2)	(6, 1, 3)	(6, 1, 4)
(6, 2, 1)	(6, 2, 2)	(6, 2, 3)	(6, 2, 4)
(6, 3, 1)	(6, 3, 2)	(6, 3, 3)	(6, 3, 4)
(6, 4, 1)	(6, 4, 2)	(6, 4, 3)	(6, 4, 4)

F4			
(4, 1, 1)	(4, 1, 2)	(4, 1, 3)	(4, 1, 4)
(4, 2, 1)	(4, 2, 2)	(4, 2, 3)	(4, 2, 4)
(4, 3, 1)	(4, 3, 2)	(4, 3, 3)	(4, 3, 4)
(4, 4, 1)	(4, 4, 2)	(4, 4, 3)	(4, 4, 4)

F3			
(3, 1, 1)	(3, 1, 2)	(3, 1, 3)	(3, 1, 4)
(3, 2, 1)	(3, 2, 2)	(3, 2, 3)	(3, 2, 4)
(3, 3, 1)	(3, 3, 2)	(3, 3, 3)	(3, 3, 4)
(3, 4, 1)	(3, 4, 2)	(3, 4, 3)	(3, 4, 4)

Fig. 3. Generation of Key Matrices.

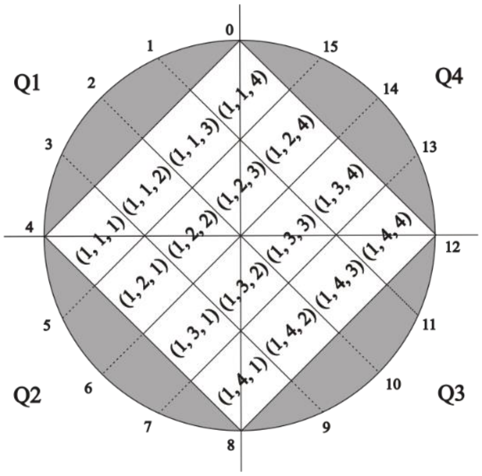


Fig. 4. Design of TCE Algorithm.

3) A secondary diagonal is defined as the collection of entries *HCs*(*f*, *i*, *j*), where $i + j = 5$ that can be calculated by the sum of the mean of the symmetric coordinates *(i, j)* and *(j, i)* of *HCs* matrix [24]. The coordinates comprise the secondary diagonal for each face (*f*) of the hybrid cube, as follows:

{(1, 4), (2, 3), (3, 2), (4, 1)}

When the value of diagonals *HCs*(*f*, *i*, *j*) satisfy the properties of primary and secondary diagonals then the value of coordinates of a particular cell is 1/2*HCs(f, i, j)*.

The quarters *Q1* to *Q4* is used to extract the coordinates value during the rotation of *HCs* and shifting the value from quarters *Q1* to *Q4* based on properties discussed in Definition 1. The value of the coordinates of *HCs* can be extracted based on the following equations in Table I.

In this technique, each quarter of the square matrix is used to generate the one coordinate for a newly generated key matrix based on the quarter's equations. Similarly, the quarters *Q1* to *Q4* are used to generate one row for the new key matrix. In this regard, four key matrices are required to generate the order 4 key matrix. Furthermore, the new key matrix is rotated from quarters *Q1* to *Q4* and the value of the coordinates are shifted according to the rotation pattern.

TABLE. I. COORDINATES EXTRACTION FROM Q1 TO Q4

Quarters	**Extraction of coordinates value**
Q1	$\sum_{i=0}^{1}\sum_{j=1+i}^{4-i}(i+1, j)$
Q2	$\sum_{j=0}^{1}\sum_{i=1+j}^{4-j}(i, j+1)$
Q3	$\sum_{i=0}^{1}\sum_{j=1+i}^{4-i}(4-i, j)$
Q4	$\sum_{j=0}^{1}\sum_{i=1+j}^{4-j}(i, 4-j)$

4) Columns and rows selection: Definition 2. Let CR_j be the determinant of columns and rows rotation, when

$$CR_j = (P_j + IV)\, mod\ m \tag{1}$$

where, CR_j is represented as the resultant value of columns and rows, denoted as ABCDWXYZ, P_j is represented as Password, IV is the initialization vector, and m is the string of 8-integer numbers.

The input of the KSAHC algorithm can be considered as N. The m and IV are considered as the pre-defined strings of an integer number and the P_j is also obtained from the user. The overall input is the addition of IV and P_j (j = 0, 1, 2, …, N) that result is modulo with respect to m which is denoted as CR_j and can be represented in Equation (2).

$$CR_j = \begin{cases} P_j & if \quad 0 \le i \le N \\ IV & if \quad j = N \in Z^+ \\ m & if \quad j = N \bmod 12345678 \end{cases} \tag{2}$$

The resultant value ABCDWXYZ of CR_j is employed for the rotation of columns and rows. The value of *ABCD* shows the rotation of columns in which the first column (C0) is rotated based on *A*, second column (C1) is rotated based on *B*, third column (C2) is rotated based on *C*, and fourth column (C3) is rotated based on *D*. Whereas, the value of WXYZ shows the rotation of rows in which first row (R0) is rotated based on W, second row (R1) is rotated based on *X*, third row (R2) is rotated based on *Y*, and fourth row (R3) is rotated based on *Z*. Also, the *IV* is a random number that is used with the password in the key schedule algorithm and it is used only one time in each session. The purpose of using *IV* is to prevent repetition in keys, which make it difficult for the cryptanalysis to find the keys pattern and break the cipher.

5) Initialization face of columns rotation: Definition 3. Let *A*, *B*, *C* and *D* be the value of hybrid cube that is obtained from the resultant value of CR_j and it is used for the initialization of column rotation. It is also defined as the first four bits of CR_j that modulo with the hybrid cube faces.

$$\text{IniFCol} = (((\text{A} + \text{B} + \text{C} + \text{D}) \bmod \text{f}) + 1) \tag{3}$$

where, *IniFCol* represent the initialization face for columns rotation and *f* represent the six faces of hybrid cube.

6) Initialization face of rows rotation: Definition 4. Let *W*, *X*, *Y* and *Z* be the value of hybrid cube that is taken from the resultant value of rows and columns selection and it is employed for the initialization of row rotation. It is also defined as the last four numbers of the value of columns and rows selection.

$$\text{IniFRow} = (((\text{W} + \text{X} + \text{Y} + \text{Z}) \bmod \text{f}) + 1) \tag{4}$$

where *IniFRow* represent the initialization face for rows rotation and *f* represent the six faces of the hybrid cube.

7) The shiftcolumns transformation: As indicated by its name, the ShiftColumns transformation processes different columns between different faces of the hybrid cube. The operation of shifting the columns of the cube states over the specified column offsets is denoted as:

ShiftColumns(States)

The 3D rotation of columns of each face of order 4 matrix with other faces of hybrid cubes depends on first four value of the CR_j. For example, the column vector C_0 face 1 matrix is shifted over the C_0 vector of face 2. The C_0 vector of face 2 is shifted over the C_0 vector of face 3, C_0 vector of face 3 over the face 4, C_0 vector of face 4 is shifted over the face 1 and the C_0 vector of face 1 is shifted over the face 2. Similarly, the columns C_1, C_2 and C_3 of each face have shifted the coordinates of the selected columns. The ShiftColumns operations can be performed on different faces of the hybrid cube, so the rotation pattern of shifting columns of different faces depends on the CR_j value.

Definition 5. Let the ShiftColumns operation be the transposition of column vectors that cyclically shifts the columns of each face over the different column offsets of the hybrid cube. If the different faces of a cube having first and fourth columns then the value of $\delta = 1$, otherwise the value of the middle column is $\delta = 0$ as shown in Equation (5).

The processing equation of ShiftColumns is computed as follows:

$$ShiftColumns = \prod_{i=1}^{4} 4Rot_i\left(C_i + \left(\delta \times Q_i\right)\right) \tag{5}$$

where Rot_i is represented as the number of rotations based on the CR_j value, Ci is the column vectors of cube faces and the Q_i is the rotation of quarters Q_1 to Q_4.

Equation (5) defines each rotation Rot_i based on the CR_j value that affects the changes of column vectors rotation into 4 times on different faces of the hybrid cube because the ShiftColumns transformation is applied on four faces of the cube. Similarly, if the CR_j value is two then it affects the column vectors 8 times, if the CR_j value is three then it affects the column vectors 12 times, if CR_j value is four then it affects the column vectors 16 times, and so on. As mentioned earlier, the rotation pattern of columns depends on the first four (*ABCD*) values of CR_j that rotates the respective columns (C_0, C_1, C_2 and C_3) shown in Fig. 5.

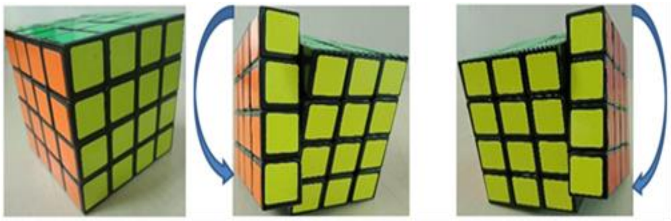


Fig. 5. ShiftColumns Transformation.

For example, the column C_0 is rotated based on the value of *A*. Suppose the value of *A* is 3 then the column vector C_0 is rotated 3 times to different faces of the cube. Similarly, the value of *B* is responsible for shifting the coordinates of respective column vector C_1, the value of *C* shifts the coordinates of column C_2 and the value of *D* shifts the coordinates of respective column C_3. Moreover, if the rotation Rot_i of the cube is having the sided columns (C_0, C_3) where the $\delta = 1$ that also affects both sides faces of the cube and it rotates using the triangular coordinate extraction technique of quarters rotation Q_i. The extraction of the coordinate's value during the rotation of *HCs* and shifting the value from quarters Q_1 to Q_4 based on the rotation pattern. The rotation of quarters depends on the left-sided column (C_0) and right-sided column (C_3) of the hybrid cube. There are two different approaches used in the quarter's rotation of hybrid cube which is clockwise and counterclockwise. Firstly, if the rotation is based on C_0 then the quarters are being rotated to counterclockwise and the number of rotations depends on the Rot_i value. So, the process of a single rotation of column is that the value of Q_1 is shifted to Q_2, Q_2 value shifted to Q_3, Q_3 value shifted to Q_4 and finally Q_4 to Q_1. Similarly, if the rotation is based on C_3 then the quarters will be rotated towards the clockwise direction and the process of a single rotation of column is that the value of Q_1 is shifted to Q_4, Q_4 is shifted to Q_3, Q_3 to Q_2 and Q_2 is finally shifted to Q_1. The number of rotation of quarters Q_i depends on the value of Rot_i in the CR_j.

For ShiftColumns transformation, various permutations are performed on the column to add confusion in the key matrices of the hybrid cube. The rotation pattern of different faces can be divided into two different cases and the pseudocode for both cases of ShiftColumns transformation used to rotate the *HCs* is shown in Algorithm 1.

a) Case 1: IniFCol value is between 1 TO 4: If the IniFCol value is between 1 to 4, then we employ the rotation pattern of column vectors of F1, F2, F3 and F4 faces. The rotation can be performed counterclockwise direction and the rotation of columns of each face affects the coordinate value of other cube faces. Each time the rotation of C0 of each face affects the rotation of quarters of face F5 and it rotates to counterclockwise using the TCE technique. Also, each time rotation of C3 in each face affects the face F6 and it rotates to clockwise using TCE technique.

b) Case 2: IniFCol value is 5 or 6: If the *IniFCol* value is 5 or 6, then the column rotation uses the *F2*, *F4*, *F5* and *F6* faces. The rotation of column 1 and column 4 affects face 3 and 1, and these faces rotate counterclockwise and clockwise, respectively with the TCE technique.

Algorithm 1. Pseudo-Code of ShiftColumn operation

```
Assign suitable values to $Rot_i$, $CR_j$
loop
Calculate ShiftColumns where case 1 and 2 are calculated using
Equation 5.
for the hybrid cube columns $C_0$ to $C_3$ do {
  if the hybrid cube column $C_0$ OR $C_3$ then
    for number of rotations $Rot_i$ based on $CR_j$ value do
      if the hybrid cube column $C_0$
        LeftFace ← RotateMatrixCounterClockWise(LeftFace)
      else
        RightFace ← RotateMatrixClockWise(RightFace)
      end if
    end for
  end if
  for number of rotations $Rot_i$ based on $CR_j$ value do
    for var i = 0 to 3 do
      $temp_{f, i, Rot} \leftarrow Col_{f, i, Rot}$
      $Col_{f, i, Rot} \leftarrow Col_{f, i, Rot}$
      $Col_{f, i, Rot} \leftarrow Col_{f, i, Rot}$
      $Col_{f, i, Rot} \leftarrow Col_{f, i, Rot}$
      $Col_{f, i, Rot} \leftarrow temp_{f, i, Rot}$
    end for
  end for
end for
```

8) The shiftrows transformation: The ShiftRows operation processes different rows between different faces of the cube. ShiftRows transformation is applied to the matrices that are generated from the ShiftColumns operation. The transformation of shifting the rows of the hybrid cube states over the specified row offsets is denoted as:

ShiftRows(States)

The number of rotations of row vectors in each face with other faces of hybrid cubes depends on last four value of the CR_j. For example, the row vector R_0 of the 4 × 4 matrix of face 1 shifted over the R_0 vector of face 2. The R_0 vector of face 5 is shifted over the R_0 vector of face 1, R_0 vector of face 1 over the face 6, R_0 of face 6 is shifted over the face 3 and the R_0 of face 3 is shifted over the face 5. Similarly, the rows R_1, R_2 and R_3 of each face is shifted based on the coordinate value of the selected rows. The ShiftRows operations can be performed on different faces of the hybrid cube, so the rotation pattern of shifting rows of different faces depends on the CR_j value.

Definition 6. Let the ShiftRows operation be a transposition of row vectors that cyclically shifts the rows of each face over different row offsets of the hybrid cube. If the different faces of a cube are having first and fourth rows then the value of $\delta = 1$, otherwise the middle row's value is $\delta = 0$ depicted in Equation (6). The mathematic formulation of ShiftRows is computed as follows:

$$ShiftRows = \prod_{i=1}^{4} 4Rot_i\left(R_i + \left(\delta \times Q_i\right)\right) \quad (6)$$

where Rot_i is represented as the number of rotations based on the CR_j value, R_i is the row vectors of cube faces and the Q_i is the rotation of quarters Q_1 to Q_4.

Each rotation of Rot_i based on the CR_j value that affects the changes of row vectors rotation into 4 times on different faces of the hybrid cube, because the ShiftRows transformation is applied on four faces of the cube and other two faces rotated clockwise, and counterclockwise based on the ShiftRows transformation. Similarly, if the CR_j value is two then it affects the rows vectors 8 times, and so on. The rotation pattern of rows depends on the last four values (*WXYZ*) of CR_j that rotates the respective rows R_0, R_1, R_2 and R_3 shown in Fig. 6.

For example, the row R_0 is rotated based on the value of *W*. Suppose the value of *W* is 3 then the row vector R_0 has rotated the coordinates into 3 times to the different faces of the cube. Similarly, the value of *X* shifts the coordinates of respective row vector R_1, the value of *Y* shifts the coordinates of row R_2 and the value of *Z* shift the coordinates of respective row R_3. Moreover, if the rotation Rot_i of the cube is having the sided rows (R_0, R_3), then the $\delta = 1$ that affects both sides faces of the cube and it rotates by using the quarter's rotation of Q_i. The rotation pattern of rows of different faces can be divided into two different cases.

a) Case 1: IniFRow value is 1, 3, 5 or 6: If the *IniFRow* value is 1, 3, 5 or 6, then the faces *F1*, *F3*, *F5* and *F6* are utilized as the rotation pattern for the row vectors. The similar process of ShiftRows rotation is performed to the cases of ShiftColumns transformation. Each time the rotation of R_0 of each face affects the face *F2* and it rotates into a counterclockwise. Similarly, the rotation of R_3 in each face affects the face *F4* and it rotates into a clockwise direction.

b) Case 2: IniFRow value is 2 or 4: If the *IniFRow* value is 2 or 4, then the faces *F2*, *F4*, *F5* and *F6* are employed for the row vectors rotation. The rotation of row 1 and row 4 affects the faces *F3* and *F1*, and these faces rotate counterclockwise and clockwise, respectively.

9) Unique matrices operation: In this section, the modulo-16 operation is applied on coordinates of all faces of hybrid cube matrices that are generated from ShiftRows transformation. Each run will give 1 value in the new modulo matrix. The modulo matrices of the hybrid cube which contains the coordinates value are in the range of 0 to 15.

Definition 7. Let the hybrid cube be the 4 × 4 matrices, if any repeated value found in the modulo matrices coordinates, then replace it using the following rules:

$a = a - 1$ for 1st repetition

$a = a - 2$ for 2nd repetition

$a = a - 3$ for 3rd repetition

It will continue until we get zero value. After reaching zero value, if repetition still exists then we will replace by using the following rules:

$a = a + 1$ for 1st repetition

$a = a + 2$ for 2nd repetition

This process will continue until we get the non-repeated matrices value.

Moreover, if the modulo matrices of the hybrid cube are consisting of repeated value(s) in each coordinate of rows and columns, then the properties of Definition 7 are applied on the newly generated modulo matrices in order to get the unique matrices value. These unique matrices will be used to calculate the value of triangular coordinate matrices based on rotation points in the next section.

10) Triangular key matrices: This section calculates the value of key matrices that are generated with ShiftRows transformation using the rotation points which are based on unique matrices. The design of rotation points of *HCs* can be divided into 4 quarters and the rotation points represented as 0 to 15 from quarter Q1 to Q4 [25]. The new key matrices are generated through the calculation of ShiftRows coordinate values based on the rotation points and finally, the value of each matrix is organized based on the unique matrices. The generation of triangular key matrices develops the confusion element in the design of key scheduling, and it increases the difficulty for the cryptanalysis to try all key possibilities. The session keys are generated from master keys by using the TCE quarters rotations that are employed to encrypt the message 1 to 4 in the HiSea encryption algorithm. The novelty of the key schedule algorithm is that all the generated master and session keys of the 3D hybrid cube are invertible and suitable for encryption and decryption in the non-binary block cipher.

B. HiSea Encryption Algorithm

The Hybrid Cube Encryption Algorithm (HiSea) is adopted as the platform in order to validate the proposed key schedule algorithm. The KSAHC algorithm is embedded with the HiSea Encryption algorithm and used to generated encryption keys to encrypt the message into ciphertext. HiSea is the symmetric non-binary block cipher because the encryption and decryption keys, plaintext, ciphertext and internal operation in the encryption or decryption process, are all based on the integer numbers [22], [26]. The Initial Matrix (IM) used during the encryption and decryption process is a secondary security measure which ensures the authenticity of the user. The plaintext is segmented into 64 characters and converted into Extended ASCII codes and the four matrices of Plaintext are represented as *P1* to *P4*. The intermediate result (*P1′*) is obtained from adding *P1* to *P4* with the *IM* and used in the encrypting process of *P2*. The intermediate result (*P2′*) is obtained from adding *P2* with *P1′* and the result is used in the encrypting process of *P3*. This process is repeated for *P4*. The major reason for integrating this method is to ensure any change made in *P1* will reflect in another ciphertext. The process of diffusion is performed using the MixCol and MixRow operations adapted from Toy100 to strengthen the ciphertext [27]. The graphical representation of the encryption process of HiSea block cipher is shown in Fig. 7.

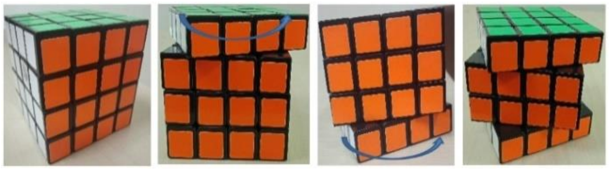


Fig. 6. ShiftRows Transformation.

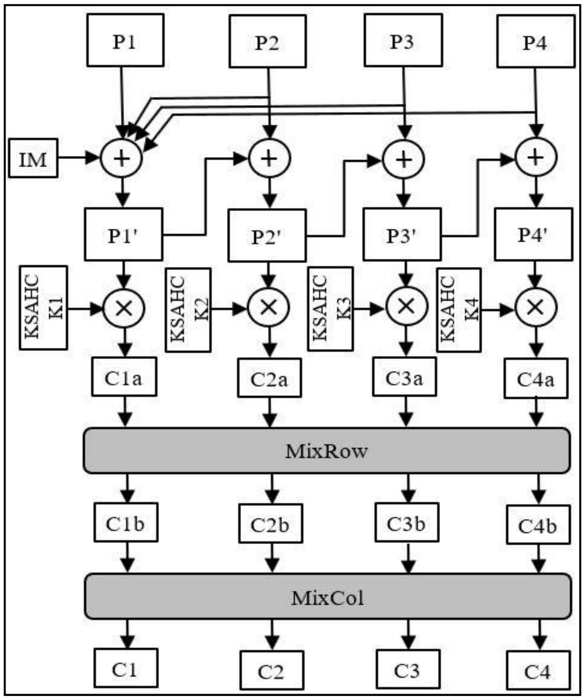


Fig. 7. Encryption Algorithm.

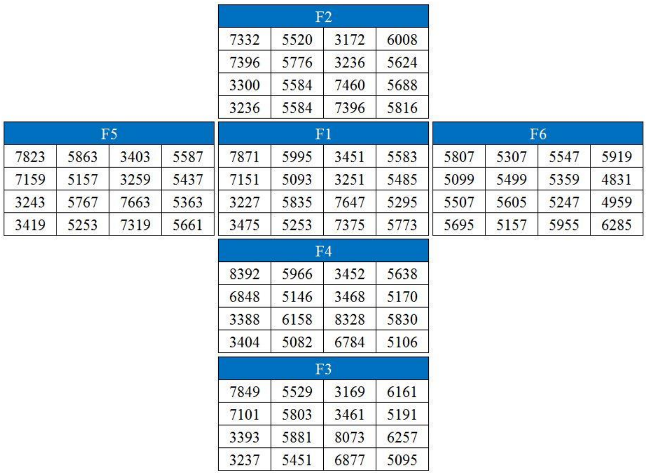
C. Decryption Algorithm

The decryption process is the reverse engineering of the encryption process in which it receives the ciphertext and secret key from the user as the requirement for reconstructing the message back to its readable form [22]. For the purpose of decryption, the recipient required the Ciphertext (C1 to C4) from sender to decrypt the ciphertext into the original readable form. The sender and receiver need to agree earlier on user password for performing the process of encryption and decryption. The password is used to generate the master key in the decryption process, all session keys are the inverse of the encryption keys *K1* to *K4*.

III. Results and Discussion

Let us consider the order 4 hybrid cubes that are used in the rotation of *HCs* which is the main element of the construction of the key scheduling algorithm. In this section, the step by step process of KSAHC algorithm with the example is described and compared with existing algorithms. Furthermore, the generated triangular key matrices have been analyzed to verify the suitability of encryption and decryption keys to the non-binary block cipher. In this regard, some experimental results include the Brute Force, entropy, correlation assessment, avalanche effect, and NIST randomness test suit has been used in the evaluation of the final output of KSAHC algorithm and cipher to prove its strength.

Firstly, the key table is generated from HiSea encryption algorithm in which 24 hybrid cube layers are used as the input to the proposed algorithm. By using the technique [24], 24 matrices are converted into six key matrices that show the six faces of the hybrid cube and these faces were used for the rotation purpose. The generated cube faces of the hybrid cube are shown in Fig. 8.



F2			
7332	5520	3172	6008
7396	5776	3236	5624
3300	5584	7460	5688
3236	5584	7396	5816

F5			
7823	5863	3403	5587
7159	5157	3259	5437
3243	5767	7663	5363
3419	5253	7319	5661

F1			
7871	5995	3451	5583
7151	5093	3251	5485
3227	5835	7647	5295
3475	5253	7375	5773

F6			
5807	5307	5547	5919
5099	5499	5359	4831
5507	5605	5247	4959
5695	5157	5955	6285

F4			
8392	5966	3452	5638
6848	5146	3468	5170
3388	6158	8328	5830
3404	5082	6784	5106

F3			
7849	5529	3169	6161
7101	5803	3461	5191
3393	5881	8073	6257
3237	5451	6877	5095

Fig. 8. Key Matrices Generated using TCE Technique.

Moreover, the string P_j is arbitrary chosen by the user as his private encryption key and is used in the columns and rows selection. Suppose, the value of CR_j is as follows:

P_j = 9876543219

IV = 96421358

m = 123456789

CR_j = (9876543219 + 96421358) mod *123456789*

= 96421457

The value of CR_j shows that the first four numbers are 9642 which are used for the column's rotation and the last four numbers 1457 is used for rows rotations. Furthermore, the value of CR_j is 96421457 that are also used for the initialization face of columns and rows rotation. In this regard, the first four numbers (9642) of CR_j are used to find the initialization face for column rotation. So, Equation (3) is used to calculate the value of *IniFCol* as defined in the following equation:

$IniFCol = ((9 + 6 + 4 + 2) \bmod 6) + 1)= 4$

The value of *IniFCol* is 4 which means the initialization face for rotation pattern in ShiftColumns transformation is face 4.

The last four numbers (1457) of CR_j is used for the initialization face of row rotation. In this regard, Equation (4) is used to calculate the value of *IniFRow* described the following equation:

$IniFRow = ((1 + 4 + 5 + 7) \bmod 6) + 1= 6$

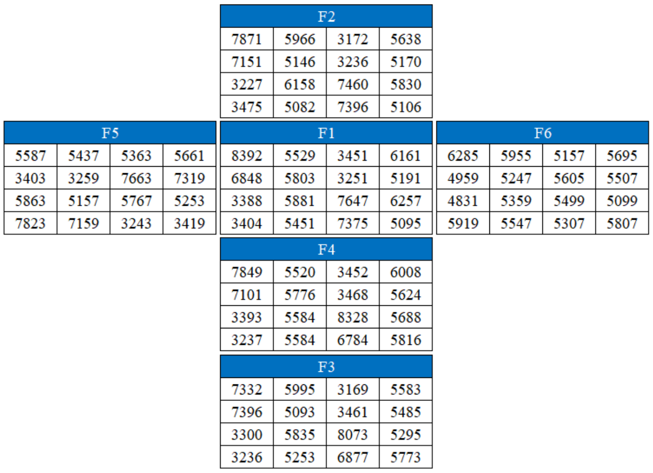
In this case, the *IniFRow* value is 6, which shows that the row rotation started in ShiftRows operation with face 6.

Based on the *IniFCol* value, the initialization face for the ShiftColumns transformation is face 4, so the rotation pattern can be made using case 1 in ShiftColumns transformation. The CR_j value for the Shifting of columns C_0 to C_3 is 9642. In this case, C_0 vectors rotated 9 times, C_1 vectors rotated 6 times, C_2 vectors rotated 4 times, and C_3 vectors rotated 2 times in the hybrid cube faces 1 to 4. The key matrices are generated using the TCE technique employed for the ShiftColumns transformation. So, case 1 is selected with faces pattern *F1*, *F2*, *F3* and *F4* for the rotation of columns. The rotation of columns C_0 and C_3 affects the other faces *F5* and *F6* that rotated counterclockwise and clockwise respectively based on the TCE quarters rotation technique. The hybrid cube key matrices after the ShiftColumns operation is presented in Fig. 9.

As discussed earlier, the *IniFRow* value for the ShiftRows operation is the face 6, so the rotation pattern can be made based on the ShiftRows case 1. The CR_j value for the Shifting of rows R_0 to R_3 is 1457. In this case, R_0 vectors are rotated 1 time, R_1 vectors rotated 4 times, R_2 vectors rotated 5 times, and R_3 vectors rotated 7 times in the hybrid cube faces 2, 4, 5 and 6. The key matrices generated using the ShiftColumns are used by ShiftRows transformation. So, we selected the faces *F1*, *F3*, *F5* and *F6* for the rotation purpose. Also, the rows R_0 and R_3 affect the other faces *F2* and *F4* that rotated counterclockwise and clockwise respectively. The hybrid cube key matrices after the ShiftRows operation is shown in Fig. 10.

The matrices generated using the ShiftRows transformation are employed for the modulo operation. The modulo matrices of the hybrid cube can be presented in Fig. 11.

The modulo matrices of the hybrid cube that contains the repeated values are depicted in Fig. 10. So, we apply the properties of Definition 7 on the hybrid cube faces in order to remove repetition and generate unique matrices. The resultant unique matrices are shown in Fig. 12. Finally, the output of ShiftRows and unique matrices with the respective coordinate's value are considered in order to calculate the triangular key matrices. The value of triangular key matrices is used according to their coordinates and then calculate the value based on rotation points [25]. The resultant six key matrices of the hybrid cube are used as the encryption and decryption key in the non-binary block cipher. The triangular key matrices are used as the master keys for the encryption process as shown in Fig. 13.



F2			
7871	5966	3172	5638
7151	5146	3236	5170
3227	6158	7460	5830
3475	5082	7396	5106

F5			
5587	5437	5363	5661
3403	3259	7663	7319
5863	5157	5767	5253
7823	7159	3243	3419

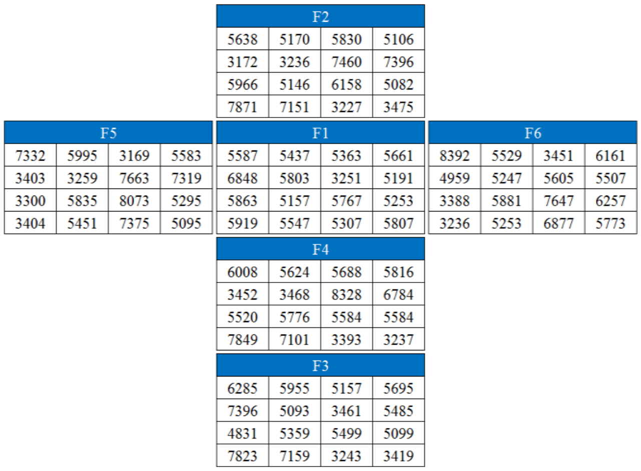
F1			
8392	5529	3451	6161
6848	5803	3251	5191
3388	5881	7647	6257
3404	5451	7375	5095

F6			
6285	5955	5157	5695
4959	5247	5605	5507
4831	5359	5499	5099
5919	5547	5307	5807

F4			
7849	5520	3452	6008
7101	5776	3468	5624
3393	5584	8328	5688
3237	5584	6784	5816

F3			
7332	5995	3169	5583
7396	5093	3461	5485
3300	5835	8073	5295
3236	5253	6877	5773

Fig. 9. Hybrid Cube Key Matrices after ShiftColumns Transformation.



F2			
5638	5170	5830	5106
3172	3236	7460	7396
5966	5146	6158	5082
7871	7151	3227	3475

F5			
7332	5995	3169	5583
3403	3259	7663	7319
3300	5835	8073	5295
3404	5451	7375	5095

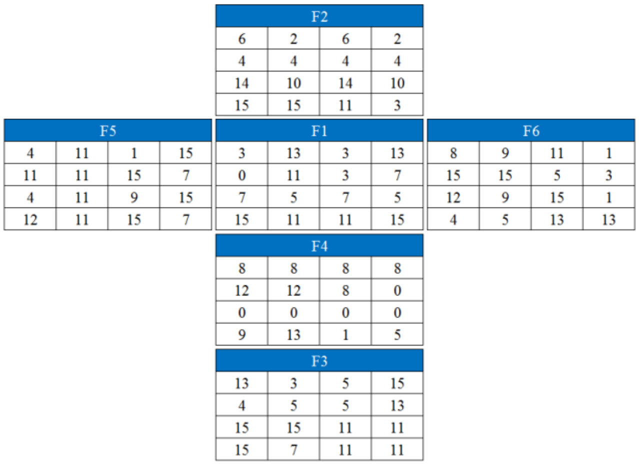
F1			
5587	5437	5363	5661
6848	5803	3251	5191
5863	5157	5767	5253
5919	5547	5307	5807

F6			
8392	5529	3451	6161
4959	5247	5605	5507
3388	5881	7647	6257
3236	5253	6877	5773

F4			
6008	5624	5688	5816
3452	3468	8328	6784
5520	5776	5584	5584
7849	7101	3393	3237

F3			
6285	5955	5157	5695
7396	5093	3461	5485
4831	5359	5499	5099
7823	7159	3243	3419

Fig. 10. Hybrid Cube Key Matrices after ShiftRows Transformation.



F2			
6	2	6	2
4	4	4	4
14	10	14	10
15	15	11	3

F5			
4	11	1	15
11	11	15	7
4	11	9	15
12	11	15	7

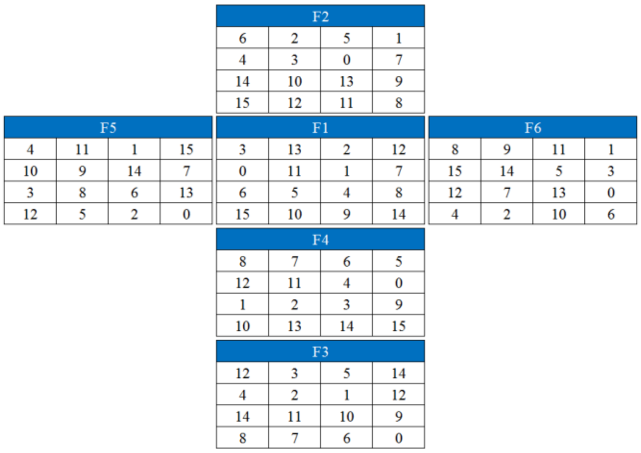
F1			
3	13	3	13
0	11	3	7
7	5	7	5
15	11	11	15

F6			
8	9	11	1
15	15	5	3
12	9	15	1
4	5	13	13

F4			
8	8	8	8
12	12	8	0
0	0	0	0
9	13	1	5

F3			
13	3	5	15
4	5	5	13
15	15	11	11
15	7	11	11

Fig. 11. Modulo-16 Matrices of the Hybrid Cube.



F2			
6	2	5	1
4	3	0	7
14	10	13	9
15	12	11	8

F5			
4	11	1	15
10	9	14	7
3	8	6	13
12	5	2	0

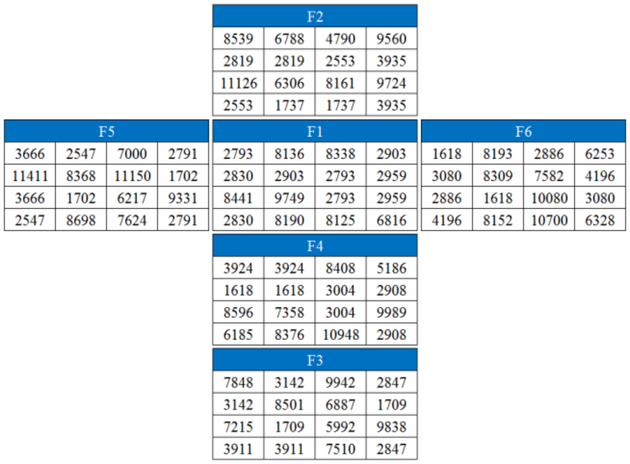
F1			
3	13	2	12
0	11	1	7
6	5	4	8
15	10	9	14

F6			
8	9	11	1
15	14	5	3
12	7	13	0
4	2	10	6

F4			
8	7	6	5
12	11	4	0
1	2	3	9
10	13	14	15

F3			
12	3	5	14
4	2	1	12
14	11	10	9
8	7	6	0

Fig. 12. Hybrid Cube Matrices using unique Matrices Operation.



F2			
8539	6788	4790	9560
2819	2819	2553	3935
11126	6306	8161	9724
2553	1737	1737	3935

F5			
3666	2547	7000	2791
11411	8368	11150	1702
3666	1702	6217	9331
2547	8698	7624	2791

F1			
2793	8136	8338	2903
2830	2903	2793	2959
8441	9749	2793	2959
2830	8190	8125	6816

F6			
1618	8193	2886	6253
3080	8309	7582	4196
2886	1618	10080	3080
4196	8152	10700	6328

F4			
3924	3924	8408	5186
1618	1618	3004	2908
8596	7358	3004	9989
6185	8376	10948	2908

F3			
7848	3142	9942	2847
3142	8501	6887	1709
7215	1709	5992	9838
3911	3911	7510	2847

Fig. 13. Triangular Key Matrices of the Hybrid Cube.

The novelty of the key schedule algorithm is that all the generated master and session keys of the hybrid cube are invertible and suitable for the encryption and decryption in the non-binary block cipher. After generating the encryption keys using KSAHC algorithm, the encryption of plaintext message with the generated key is performed using the HiSea encryption algorithm [22]. The generated keys have been tested with the existing block cipher (HiSea) and the steps of encryption are demonstrated by using the following message "Hybrid Cubes Encryption Algorithm is originated from UTHM JOHOR." and user password "9876543219". The input message from the user is converted into 64 Extended ASCII codes and the four matrices are represented into 4 × 4 matrices. These matrices are shown as follows:

$$P1 = \begin{bmatrix} 72 & 121 & 98 & 114 \\ 105 & 100 & 32 & 67 \\ 117 & 98 & 101 & 115 \\ 32 & 69 & 110 & 99 \end{bmatrix}, \quad P2 = \begin{bmatrix} 114 & 121 & 112 & 116 \\ 105 & 111 & 110 & 32 \\ 65 & 108 & 103 & 111 \\ 114 & 105 & 116 & 104 \end{bmatrix}$$

$$P3 = \begin{bmatrix} 109 & 32 & 105 & 115 \\ 32 & 111 & 114 & 105 \\ 103 & 105 & 110 & 97 \\ 116 & 101 & 100 & 32 \end{bmatrix}, \quad P4 = \begin{bmatrix} 102 & 114 & 111 & 109 \\ 32 & 85 & 84 & 72 \\ 77 & 32 & 74 & 79 \\ 72 & 79 & 82 & 46 \end{bmatrix}$$

The initial matrix between the sender and receiver is set as follows:

$$IM = \begin{bmatrix} 540 & 3534 & 1872 & 10 \\ 24 & 1710 & 3780 & 442 \\ 3294 & 456 & 10 & 2068 \\ 1886 & 44 & 378 & 3520 \end{bmatrix}$$

The intermediate result (*P1´*) is generated by mixing the *P1*, *P2*, *P3*, *P4* and *IM* as given in Fig. 7. The value of *P1´* is shown as follows:

$$P1' = \begin{bmatrix} 937 & 3922 & 2298 & 464 \\ 298 & 2117 & 4120 & 718 \\ 3656 & 799 & 398 & 2470 \\ 2220 & 398 & 786 & 3801 \end{bmatrix}$$

Session keys are generated using the master key *F2*, the results are as follows:

$$K1 = \begin{bmatrix} 9560 & 3935 & 9724 & 3935 \\ 4790 & 2553 & 8161 & 1737 \\ 6788 & 2819 & 6306 & 1737 \\ 8539 & 2819 & 11126 & 2553 \end{bmatrix} \quad K2 = \begin{bmatrix} 3935 & 1737 & 1737 & 2553 \\ 9724 & 8161 & 6306 & 11126 \\ 3935 & 2553 & 2819 & 2819 \\ 9560 & 4790 & 6788 & 8539 \end{bmatrix}$$

$$K3 = \begin{bmatrix} 2553 & 11126 & 2819 & 8539 \\ 1737 & 6306 & 2819 & 6788 \\ 1737 & 8161 & 2553 & 4790 \\ 3935 & 9724 & 3935 & 9560 \end{bmatrix} \quad K4 = \begin{bmatrix} 8539 & 6788 & 4790 & 9560 \\ 2819 & 2819 & 2553 & 3935 \\ 11126 & 6306 & 8161 & 9724 \\ 2553 & 1737 & 1737 & 3935 \end{bmatrix}$$

The intermediate result of *P1´* is then mixed with *K1*. The matrix *C1a* is given as follows:

$$C1a = \begin{bmatrix} 47305020 & 21486039 & 60772482 & 15675827 \\ 47086872 & 20215653 & 54143777 & 13839353 \\ 62571524 & 24511099 & 72062591 & 22771459 \\ 60921727 & 22682547 & 72081800 & 20496261 \end{bmatrix}$$

After that, MixRow function is applied on *C1a*, the matrix *C1b* is mentioned as follows:

$$C1b = \begin{bmatrix} 84466886 & 129563541 & 97934348 & 123753329 \\ 81141878 & 121446302 & 88198783 & 115070002 \\ 109854082 & 159145214 & 119345149 & 157405574 \\ 104100535 & 155686074 & 115260608 & 153499788 \end{bmatrix}$$

Furthermore, the MixCol function is applied on *C1b* and generates Ciphertext *C1* from Plaintext message P1 is given as follows:

$$C1 = \begin{bmatrix} 269709299 & 406695917 & 301393739 & 392323119 \\ 275462846 & 410155057 & 305478280 & 396228905 \\ 295096495 & 436277590 & 322804540 & 425975364 \\ 298421503 & 444394829 & 332540105 & 434658691 \end{bmatrix}$$

Similarly, the Plaintext message *P2*, *P3* and *P4* follows the similar process as for *P1* and generate the final ciphertext *C2*, *C3* and *C4*. The Ciphertext (*C2* to *C4*) are written as follows:

$$C2 = \begin{bmatrix} 417694123 & 372785950 & 363286275 & 417594055 \\ 406785499 & 361116578 & 352065797 & 403183993 \\ 376710203 & 341235079 & 323736215 & 385917619 \\ 413865098 & 369410576 & 355765496 & 418881909 \end{bmatrix}$$

$$C3 = \begin{bmatrix} 436820789 & 337379515 & 451904811 & 302898477 \\ 438423640 & 337460276 & 453995546 & 301145893 \\ 468659513 & 359242079 & 479584329 & 322486079 \\ 479934787 & 364471642 & 491438341 & 332868618 \end{bmatrix}$$

$$C4 = \begin{bmatrix} 443632423 & 387993015 & 396962894 & 448294179 \\ 460488355 & 401090600 & 410736380 & 461558698 \\ 442555760 & 377415801 & 390990607 & 439180809 \\ 429290483 & 361006024 & 379718683 & 420514922 \end{bmatrix}$$

Based on the obtained results, the key schedule algorithm and the ciphertext *C1* to *C4* have been evaluated and tested. The results are described in the following sub-sections.

A. Brute Force Attack

In general, this attack is possible if an adversary could generate one possible correct key from a large key space. The attacker has no knowledge of encryption key(s), so the attacker generates and computes every possible combination of the encryption key to recover the secret key that was used for the encryption process. In order to achieve the optimum security level, the key space must be at least 2^{128} to resist the Brute Force attack [28]. In this cipher, the KSAHC encryption keys are represented as $n \times n \times n$ matrix of integer numbers and used in the development of the permutation and substitution of order 4 square matrix. Each entry of encryption key lies between the range of from 1618 to 11198 or within 2^{14} bits (approx.). So, the key space for encryption and decryption keys calculated as follows:

$$2^{14} \times 2^{14} \times 2^{14} \times \ldots\ldots 2^{14} = (2^{14})^{16} = 2^{224}$$

or approximately the number of alternative keys

$$= 2.70 \times 10^{67} \text{ keys}$$

Furthermore, the comparison of AES, DKSA, HiSea and the proposed algorithm based on the Brute Force has been calculated and presented in Table II. The number of alternative keys shows that the KSAHC algorithm is computationally secured and has a large key space which makes the brute-force attack difficult and time-consuming. Hence, the number of keys used in the HiSea with KSAHC algorithm can determine the practically infeasible to conduct a brute-force attack due to the limitation of computational power and length of the time.

B. Entropy

In this test, the strength of overall implementation of KSAHC is estimated by using random matrix technique. The strength of the master key matrices of proposed KSAHC algorithm is calculated by using MATLAB function CalculateEnt() and compared with the HiSea Key Schedule Algorithm (KSA). The KSAHC triangular key matrices shown in Fig. 13 are used to estimate the normalized Shannon entropy.

The average normalized Shannon entropy for the HiSea KSA matrices are 0.8491 and the proposed KSAHC triangular matrices are 0.9466 as all the entropy results should be closer to 1 rather than 0. The result shows that the strength of proposed KSAHC triangular key matrices is better than the HiSea KSA as shown in Table III. Hence, each matrix of triangular key matrices that represents the hybrid cube blocks consist of 16 decimal numbers, is average of 94.66% random which can be considered as almost random and it is suitable for the development of non-binary block cipher.

The result obtained from the entropy test has been compared with its AES, HiSea and DKSA counterparts. For the purpose of comparison between these block cipher, four different ciphertexts from slightly different input keys were generated and the encryption process performed with a similar message with each of the generated keys.

Based on the results from Fig. 14, the average entropy of AES has 0.9273, HiSea has 0.9830, DKSA has 0.9367, while the proposed cipher has higher average entropy of 0.9968. The results show that the generated ciphertext produces highly random output that makes it difficult for the cryptanalyst to observe the behavior and changes on the output for the purpose of attack as all the outputs are different and did not reveal any relationship between one another.

C. Correlation Assessment

The correlation test has been conducted between the blocks of encryption keys (*F1* to *F6*) and their session keys generated using KSAHC algorithm, to figure out if any predictable pattern exists among them. The value of predictable pattern between the encryption keys and session keys should be closer to zero rather than 1, because if the result is closer to 1 then it is easy for a cryptanalyst to predict other keys due to more similarities. In this test, two sets of keys are required for the testing purpose and all 4 session keys are employed to determine whether the similarity exists among the keys or not. The four-session keys of all faces (*F1* to *F6*) of the hybrid cube can be represented as *S_K1*, *S_K2*, *S_K3* and *S_K4* shown in Table IV. The average correlation between all session keys of different faces of the 3D hybrid cube appears as -0.009472 which is closer to 0 and that means there is no correlation exist between the session keys, thus makes the related key attack very difficult.

TABLE. II. COMPARISON OF KEY SPACES BASED ON THE BRUTE FORCE

Key size (bits)	Algorithm	No. of Alternative Keys	Time required at 10^9 decryption (years)	Time required at 10^{13} decryption (years)
128	AES	2^{128}= 3.4 × 10^{38}	2^{127} ns = 5.3 × 10^{21}	5.3 × 10^{17}
128	DKSA	2^{128}= 3.4 × 10^{38}	2^{127} ns = 5.3 × 10^{21}	5.3 × 10^{17}
192	HiSea	2^{192} = 6.3 × 10^{57}	2^{191} ns = 9.8 × 10^{40}	9.8 × 10^{36}
224	KSAHC	2^{224}= 2.70 × 10^{67}	2^{223} ns = 8.6 × 10^{49}	8.6 × 10^{47}

TABLE. III. NORMALIZE SHANNON ENTROPY OF PROPOSED KSAHC ALGORITHM

Keys	HiSea KSA	Proposed KSAHC
K1	0.8448	0.9514
K2	0.8644	0.9426
K3	0.8430	0.9520
K4	0.8476	0.9463
K5	0.8473	0.9385
K6	0.8473	0.9492
Average	0.8491	0.9466

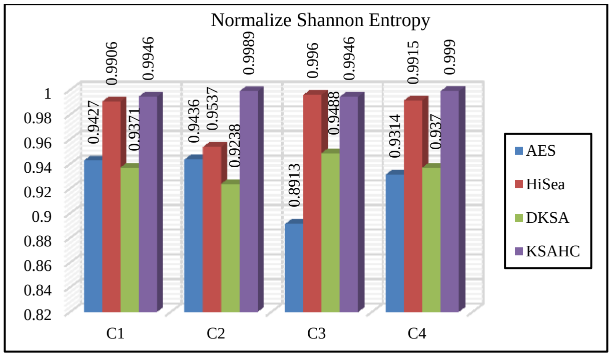


Fig. 14. Entropy Test with the different Encryption Algorithm.

In Table V, the proposed KSAHC algorithm has an average correlation assessment of -0.000601 while it ultimately compared with the AES, HiSea and DKSA, they have the average correlation of 0.185622, -0.0779 and -0.0419, respectively. The proposed KSAHC has outperformed the most widely used AES algorithm in terms of correlation. It is concluded that all the session keys are individually generated and there exist no relationship with each other. It is difficult for cryptanalyst to conduct a related key attack, even if the cryptanalyst manages to get one session key, but the other keys are unrelated and independent, so it is not easy to get all keys. The graphical representation of the correlation test between the different algorithms is shown in Fig. 15.

TABLE. IV. COMPARISON OF SESSION KEYS OF DIFFERENT FACES OF THE HYBRID CUBE

Faces	**x**	**y**	**Correlation**		**Faces**	**x**	**y**	**Correlation**
F1	S_K1_F1	S_K2_F1	-0.191778		F4	S_K1_F4	S_K2_F4	0.146547
	S_K1_F1	S_K3_F1	-0.362025			S_K1_F4	S_K3_F4	-0.103435
	S_K1_F1	S_K4_F1	0.535537			S_K1_F4	S_K4_F4	0.147765
	S_K2_F1	S_K3_F1	0.111523			S_K2_F4	S_K3_F4	0.147007
	S_K2_F1	S_K4_F1	-0.117235			S_K2_F4	S_K4_F4	-0.138630
	S_K3_F1	S_K4_F1	-0.0082772			S_K3_F4	S_K4_F4	-0.148100
F2	S_K1_F2	S_K2_F2	-0.022596		F5	S_K1_F5	S_K2_F5	0.140618
	S_K1_F2	S_K3_F2	0.127745			S_K1_F5	S_K3_F5	-0.055084
	S_K1_F2	S_K4_F2	-0.092219			S_K1_F5	S_K4_F5	-0.149065
	S_K2_F2	S_K3_F2	-0.022596			S_K2_F5	S_K3_F5	-0.117827
	S_K2_F2	S_K4_F2	0.070664			S_K2_F5	S_K4_F5	0.035217
	S_K3_F2	S_K4_F2	-0.076288			S_K3_F5	S_K4_F5	0.296118
F3	S_K1_F3	S_K2_F3	-0.081496		F6	S_K1_F6	S_K2_F6	-0.356507
	S_K1_F3	S_K3_F3	-0.272481			S_K1_F6	S_K3_F6	0.221452
	S_K1_F3	S_K4_F3	-0.081801			S_K1_F6	S_K4_F6	-0.326664
	S_K2_F3	S_K3_F3	-0.08109			S_K2_F6	S_K3_F6	-0.356356
	S_K2_F3	S_K4_F3	0.272784			S_K2_F6	S_K4_F6	0.193601
	S_K3_F3	S_K4_F3	-0.081034			S_K3_F6	S_K4_F6	0.455002
The average correlation between the face F1 to F6								-0.009472

TABLE. V. COMPARISONS OF DIFFERENT ALGORITHMS BASED ON CORRELATION

x	**y**	**AES**	**HiSea**	**DKSA**	**Proposed KSAHC**
S_K1	S_K2	0.501405	0.52327	0.041818	0.359631
S_K1	S_K3	0.389243	-0.65528	-0.42882	0.207383
S_K1	S_K4	0.353688	-0.85956	0.123036	-0.361394
S_K2	S_K3	-0.22499	0.979662	0.076945	0.008391
S_K2	S_K4	0.05817	0.460786	0.173298	-0.063184
S_K3	S_K4	0.036223	-0.91629	-0.23768	-0.154434
Average Correlation		0.185622	-0.0779	-0.0419	-0.000601

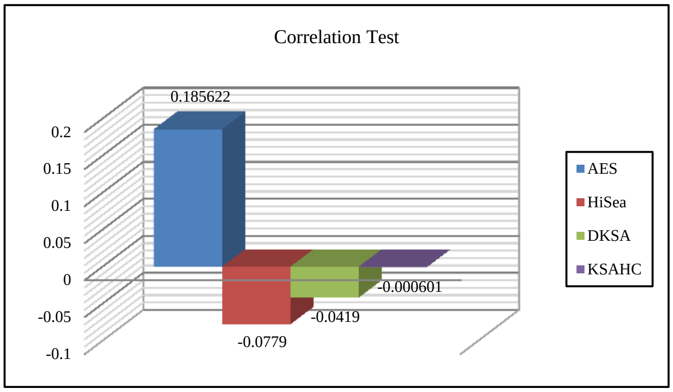


Fig. 15. Comparative Analysis of Average Correlation of different Algorithms.

D. Avalanche Effect

This test describes the behavior of the algorithm which determines the slight changes in the input that significantly affects the output value. In other words, the avalanche effect is used to measure the dissimilarity between the input and output changes. If block cipher exhibits ineffective avalanche property, the output would not be random and independently generated and the cryptanalyst could easily exploit and predict the input from the output. An efficient avalanche property of a cryptographic algorithm should be greater than or equal to 50% (≥ 50%) [29].

In this test, several input strings of our proposed algorithm have been tested to verify how much it affects the output by changing the beginning, middle and end of the inputs as shown in Table VI. The avalanche test of the proposed algorithm with different inputs produced an average result of 93% that means the proposed algorithm has favorable avalanche effect compared to others.

The avalanche effect of the session keys of different faces of the hybrid cube was tested and the proposed algorithm produces entirely different session keys in each rotation. As we mentioned earlier, two set of session keys used to calculate the avalanche effect of face *F1* to *F6* to examine the difference in the output. Also, the four-session keys from each face were compared with their counterpart session keys of different inputs, all session keys appear to be non-linear and different from each other. Hence, the average correlation of face *F1* to *F6* shown in Table VII proves that the proposed algorithm has stronger avalanche property.

A comparison of four cryptographic algorithms (AES, HiSea, DKSA and proposed KSAHC) based on different set of round keys or session keys were generated and tested to observe the changes in each session key. Based on the test, DKSA appears as the lower avalanche test score of 88% while the AES and HiSea having the avalanche score of 90%. On the other hand, the proposed KSAHC achieved the highest score in avalanche test that is 93% which means that the proposed KSAHC algorithm always produces a different set of keys for every small change in the input.

TABLE. VI. AVALANCHE EFFECT OF PROPOSED ALGORITHM WITH DIFFERENT INPUTS

Inputs	Output Keys	Aval.
0876543219	2793 8136 7066 1709 1709 6126 2830 3911 8441 5032 2793 9194 11150 2830 9737 3911	94 %
8876543219	8200 1618 4794 8056 2819 9697 3924 1618 2553 9679 3924 8075 2819 2553 7959 7249	
9876543219	1618 8193 2886 6253 3080 8309 7582 4196 2886 1618 10080 3080 4196 8152 10700 6328	91 %
9886543219	3666 3142 11150 2959 3666 7959 1618 6460 8056 10145 8790 3142 5032 2959 1618 6102	
9876543219	3004 2908 11206 1737 9194 8136 10002 8896 1737 6126 6900 3935 5181 3935 2908 3004	95%
9876653219	2886 8170 6788 1618 2791 9974 2791 8063 1618 6913 6310 2886 9647 3924 8541 3924	
9876653219	1618 2553 2553 1618 11560 4196 3004 11198 8055 5076 6190 4196 6932 7275 3004 9852	94%
9876653210	3935 1737 1737 3935 6144 4196 3004 11198 9697 5076 8541 4196 8170 6932 7275 3004	
1234567890	2959 8487 2903 5181 4196 9942 4196 10080 2903 2959 3080 3080 7215 9130 6328 8075	94%
1234567891	7848 3142 7582 2847 8309 2903 3142 8578 2903 8056 8226 2959 2959 7510 7959 2847	
2234567891	3912 1710 7067 2831 2794 10077 7849 3912 2794 8187 7511 9119 1710 8043 6306 2831	89%
3234567891	2887 2887 2792 4196 1702 6306 11199 8171 4196 5077 8542 2792 7111 9395 9119 1702	
Average Avalanche		93%

TABLE. VII. AVALANCHE EFFECT OF DIFFERENT FACES OF THE HYBRID CUBE

Faces	Key and Session Key		Avalanche		Faces	Key and Session Key		Avalanche
F1	S_K1_F1	S_K2_F1	94%		F4	S_K1_F4	S_K2_F4	92%
	S_K1_F1	S_K3_F1	94%			S_K1_F4	S_K3_F4	95%
	S_K1_F1	S_K4_F1	94%			S_K1_F4	S_K4_F4	92%
	S_K2_F1	S_K3_F1	93%			S_K2_F4	S_K3_F4	94%
	S_K2_F1	S_K4_F1	97%			S_K2_F4	S_K4_F4	89%
	S_K3_F1	S_K4_F	94%			S_K3_F4	S_K4_F4	89%
F2	S_K1_F2	S_K2_F2	91%		F5	S_K1_F5	S_K2_F5	91%
	S_K1_F2	S_K3_F2	85%			S_K1_F5	S_K3_F5	97%
	S_K1_F2	S_K4_F2	94%			S_K1_F5	S_K4_F5	92%
	S_K2_F2	S_K3_F2	94%			S_K2_F5	S_K3_F5	94%
	S_K2_F2	S_K4_F2	86%			S_K2_F5	S_K4_F5	94%
	S_K3_F2	S_K4_F2	92%			S_K3_F5	S_K4_F5	91%
F3	S_K1_F3	S_K2_F3	97%		F6	S_K1_F6	S_K2_F6	95%
	S_K1_F3	S_K3_F3	91%			S_K1_F6	S_K3_F6	92%
	S_K1_F3	S_K4_F3	92%			S_K1_F6	S_K4_F6	97%
	S_K2_F3	S_K3_F3	91%			S_K2_F6	S_K3_F6	98%
	S_K2_F3	S_K4_F3	95%			S_K2_F6	S_K4_F6	94%
	S_K3_F3	S_K4_F3	97%			S_K3_F6	S_K4_F6	95%
Average avalanche test between the face F1 to F6								93%

Moreover, the promising results show that the related key attack and ciphertext-only attack will be extremely difficult or even impossible. The KSAHC algorithm shows a very good avalanche property as compared with the existing algorithms as shown in Fig. 16.

E. The NIST Test

In order to analyze the randomness of the proposed scheme, the statistical test suite developed by the NIST is used. The purpose of the NIST test suite is to determine the randomness of a sequence. Any cryptographic algorithm is considered to pass the NIST test, if the resulting P-value is greater than 0.01 then it is said to be the random [30]. For that purpose, three statistical tests (frequency mono-bit test, block frequency test, and runs test) have been conducted on the output of the proposed scheme. Based on Table VIII, the results show that the P-value for the outputs of the keys of six faces of the Hybrid cube is greater than 0.01. Hence, it can be concluded that the results are in favor of the proposed algorithm and the sequence generated by the proposed KSAHC algorithm are random.

Similarly, the comparison of AES, HiSea, DKSA and proposed KSAHC has been conducted based on the NIST test to figure out the randomness of the proposed scheme as compared to other cryptographic algorithms. Table IX shows that the comparison of proposed KSAHC with different cryptographic algorithms based on the frequency test has outperformed AES, HiSea, and DKSA in term of randomness. Meanwhile, Table X shows that the proposed algorithm has achieved a better result as compared to AES and DKSA but the HiSea appears to have outperformed the proposed algorithm in terms of block frequency test. Also, Table XI shows that the comparison based on the NIST runs test in which the proposed algorithm performed better compared to HiSea but the DKSA and AES appears to have better results compared to the proposed algorithm.

The proposed KSAHC algorithm shows a better result in frequency and block frequency test as compared to AES and DKSA but the HiSea shows the better result in the block frequency test. While the results of the proposed scheme underperformed in the Runs test.

Fig. 17 shows that the average results of the frequency test are 0.6935, 0.5921 in frequency block test and 0.6486 in the run test. So, the requirement of the pseudorandom number generator has been achieved by the proposed algorithm. Hence, the results of diffusion test are in favor of the proposed algorithm and the generated sequence has an efficient diffusion property.

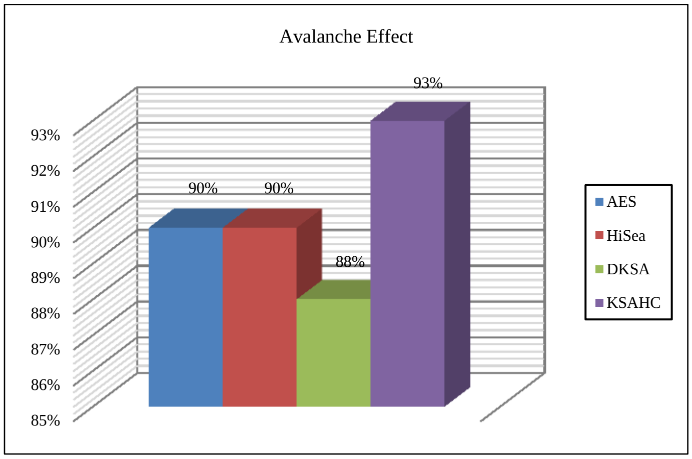


Fig. 16. The Average Result of Avalanche Test on Four different Algorithms.

TABLE. VIII. NIST TEST ANALYSIS OF PROPOSED ALGORITHM

Faces Key ID	**Frequency Test**	**Block Frequency Test**	**Runs Test**		**Faces Key ID**	**Frequency Test**	**Block Frequency Test**	**Runs Test**
S_K1_F1	0.2485	0.5745	0.1791		S_K1_F4	0.8918	0.9181	0.7845
S_K2_F1	0.1559	0.1266	0.7324		S_K2_F4	0.2185	0.1621	0.9733
S_K3_F1	0.4142	0.1660	0.0621		S_K3_F4	0.1950	0.3585	0.4629
S_K4_F1	0.5862	0.6850	0.5719		S_K4_F4	0.0656	0.7673	0.8025
S_K1_F2	0.6299	0.5474	0.8237		S_K1_F5	0.7331	0.4631	0.8316
S_K2_F2	0.2164	0.1350	0.8642		S_K2_F5	0.2463	0.6329	0.3969
S_K3_F2	0.1483	0.0748	0.0316		S_K3_F5	0.5862	0.7863	0.8008
S_K4_F2	0.7319	0.3585	0.4460		S_K4_F5	0.9454	0.1229	0.3042
S_K1_F3	0.6817	0.6045	0.9908		S_K1_F6	0.6817	0.4631	0.9908
S_K2_F3	0.4962	0.7623	0.0129		S_K2_F6	0.8379	0.6887	0.2474
S_K3_F3	0.2463	0.2283	0.8019		S_K3_F6	0.4163	0.6121	0.7210
S_K4_F3	0.9456	0.1229	0.8376		S_K4_F6	0.8379	0.6045	0.6350
Mean	0.4584	0.3655	0.5295		Mean	0.5546	0.5483	0.6626
Average NIST test between the face F1 to F6						0.5065	0.4569	0.5960

TABLE. IX. COMPARISON OF THE DIFFERENT ALGORITHMS BASED ON NIST FREQUENCY TEST

Key ID	**AES**	**HiSea**	**DKSA**	**Proposed KSAHC**
K1_1	0.3768	0.7022	0.6171	0.6817
K1_2	0.0205	0.5924	0.6171	0.8379
K1_3	1.0000	0.4913	0.4386	0.4163
K1_4	0.5959	0.9390	1.0000	0.8379
Average	0.4983	0.6812	0.6682	0.6935

TABLE. X. COMPARISON OF THE DIFFERENT ALGORITHMS BASED ON BLOCK FREQUENCY TEST

Key ID	**AES**	**HiSea**	**DKSA**	**Proposed KSAHC**
K1_1	0.7776	0.5213	0.1512	0.4631
K1_2	0.3189	0.9396	0.2017	0.6887
K1_3	0.6728	0.7409	0.2906	0.6121
K1_4	0.4884	0.4898	0.4335	0.6045
Average	0.5644	0.6729	0.2693	0.5921

TABLE. XI. COMPARISON OF THE DIFFERENT ALGORITHMS BASED ON NIST RUNS TEST

Key ID	**AES**	**HiSea**	**DKSA**	**Proposed KSAHC**
K1_1	0.4123	0.9479	0.9750	0.9908
K1_2	0.4931	0.9564	0.9750	0.2474
K1_3	1.0000	0.4675	0.2621	0.7210
K1_4	0.8790	0.1461	0.6171	0.6350
Average	0.6961	0.6295	0.7073	0.6486

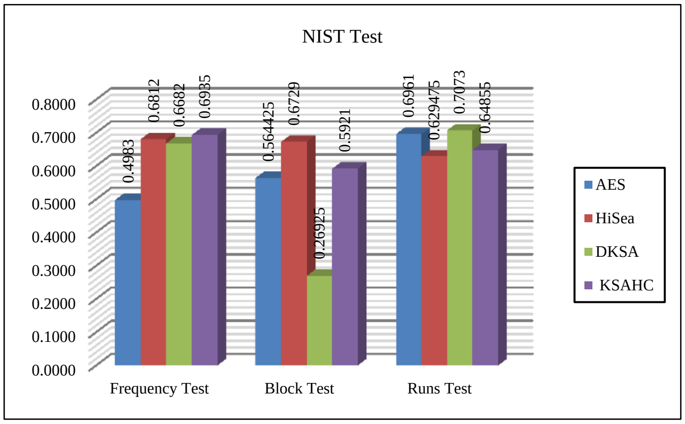


Fig. 17. Comparison of the different Algorithms based on NIST Test.

IV. CONCLUSION AND FUTURE WORK

In this paper, the KSAHC algorithm based on TCE technique is presented that is used to generate the encryption and decryption keys for the non-binary block cipher. The permutation and combination of the 3D hybrid cube from the set of integers, triangular coordinate extraction technique, and rotation of HCs are used in the design of KSAHC algorithm and it is suitable for HiSea encryption algorithm. The Brute Force and entropy test were carried out to demonstrate the strength of the keys which is highly random and having the large key space that makes it difficult and time-consuming for predicting any key pattern. The average result of the correlation is -0.000601 that closer to zero which shows no correlation exists between the input and output. Also, the result of avalanche effect is 93% which means that the proposed algorithm always produces a different set of keys for every small change in the input and it makes the attack extremely difficult or even impossible on the proposed KSAHC algorithm. Furthermore, the NIST test used to analyze the randomness of the sequences generated by the proposed scheme. The frequency mono-bit test, block frequency test and runs test has been conducted and the result of *P*-value obtained is greater than 0.01. Hence, it can be concluded that the results from the diffusion test are in favor of the proposed algorithm and the sequence generated by the proposed KSAHC algorithm has an efficient diffusion property. The results obtained from this analysis are employed to improve the overall design of the HiSea encryption algorithm. In future work, the proposed algorithm for non-binary block cipher produces only 128 bits keys but it can be upgraded into 256 and 512-bit keys that will enhance the security and performance of the algorithm. Furthermore, the proposed algorithm will be upgraded into Authenticated Encryption because of its outstanding performance.

ACKNOWLEDGMENT

The authors would like to thank the Universiti Tun Hussein Onn Malaysia and Ministry of Higher Education (MOHE) Malaysia for supporting this research under Fundamental Research Grant Scheme (FRGS), Vote No. 1642.

REFERENCES

[1] M. Ebrahim, S. Khan, and U. Bin Khalid, “Symmetric Algorithm Survey: A Comparative Analysis,” Int. J. Comput. Appl., vol. 61, no. 20, pp. 12–19, 2013.

[2] L. Savu, “Cryptography Role in Information Security,” Recent Res. Commun. Inf. Technol., pp. 36–41, 2011.

[3] A. H. Disina, Z. A. Pindar, and S. Jamel, “Enhanced Caeser Cipher to Exclude Repetition and Withstand Frequency Cryptanalysis,” J. Netw. Inf. Secur., 2015.

[4] J. Daemen and V. Rijmen, The Design of Rijndael - The Advanced Encryption Standard. 2002.

[5] M. F. Mushtaq, S. Jamel, A. H. Disina, Z. A. Pindar, N. S. A. Shakir, and M. M. Deris, “A Comprehensive Survey on the Cryptographic Encryption Algorithms,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 11, pp. 333–344, 2017.

[6] M. F. Mushtaq, U. Akram, I. Khan, S. N. Khan, A. Shahzad, and A. Ullah, “Cloud Computing Environment and Security Challenges: A Review,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 10, pp. 183–195, 2017.

[7] S. Jamel, T. Herawan, and M. M. Deris, “A cryptographic algorithm based on hybrid cubes,” Comput. Sci. Its Appl. ICCSA, vol. 6019, pp. 175–187, 2010.

[8] A. H. Disina, S. Jamel, M. Aamir, Z. A. Pindar, M. M. Deris, and K. M. Mohamad, “A Key Scheduling Algorithm Based on Dynamic Quasigroup String Transformation and All-Or- Nothing Key Derivation Function,” J. Telecommun. Electron. Comput. Eng., vol. 9, no. 3–5, pp. 1–6, 2017.

[9] N. I. of S. NIST, Advanced Encryption Standard (AES). 2001.

[10] Q.-A. Kester, “A Hybrid Cryptosystem Based on Vigenere Cipher and Columnar Transposition Cipher,” Int. J. Adv. Technol. Eng. Res., vol. 3, no. 1, pp. 141–147, 2013.

[11] R. Hoda, “Finding the total number of legal permutations of the Rubik ’ s Cube,” in Extended Essay–Mathematics, Trondheim Katedralskole, 2010, pp. 1–32.

[12] G. Hanchinamani and L. Kulakarni, “A new approach for image encryption based on cyclic rotations and multiple blockwise diffusions using Pomeau-manneville and sin maps,” J. Comput. Sci. Eng., vol. 8, no. 4, pp. 187–198, 2014.

[13] B. Nini and D. Bouteldja, “Virtual Cylindrical View of a Color Image for its Permutation for an Encryption Purpose,” Int. J. Comput. Appl., vol. 16, no. 1, pp. 11–17, 2011.

[14] J. Shen, X. Jin, and C. Zhou, “A color image encryption algorithm based on magic cube transformation and modular arithmetic operation,” Adv. Multimed. Inf. Process., vol. 3768, pp. 270–280, 2005.

[15] N. Anghel, “Determinant Identities and the Geometry of Lines and Circles,” Analele Stiint. Ovidius Constanta, Versita, vol. 22, no. 2, pp. 37–49, 2014.

[16] D. Hearn and M. P. Baker, Computer Graphics - C version. Pearson Education, 2005.

[17] W. Kuhnel, Differential Geometry, Third Edit. American Mathematical Society, 2015.

[18] M. Trenkler, “An algorithm for making magic cubes,” Pi ME J., vol. 12, no. 2, pp. 105–106, 2005.

[19] A. V. Diaconu and K. Loukhaoukha, “An improved secure image encryption algorithm based on rubik’s cube principle and digital chaotic cipher,” Math. Probl. Eng., pp. 1–10, 2013.

[20] A. B. Abugharsa, A. S. B. H. Basari, and H. M. Almangush, “A New Image Scrambling Technique using Block Rotation Algorithm based on Rubik ’ s Cube,” Aust. J. Basic Appl. Sci., vol. 7, no. 14, pp. 97–108, 2014.

[21] D. Rajavel and S. Shantharajah, “Scrambling algorithm for encryption of text using cube rotation artificial intelligence technique,” Biomed. Res., pp. 251–256, 2016.

[22] S. Jamel, M. M. Deris, I. T. R. Yanto, and T. Herawan, “The hybrid cubes encryption algorithm (HiSea),” Commun. Comput. Inf. Sci. Springer-Verlag Berlin Heidelb., vol. 154, pp. 191–200, 2011.

[23] D. Rajavel and S. P. Shantharajah, “Cryptography Based on Combination of Hybridization and Cube ’ s Rotation,” Int. J. Comput. Intell. Informatics, vol. 1, no. 4, pp. 294–299, 2012.

[24] M. F. Mushtaq, S. Jamel, and M. M. Deris, “Triangular Coordinate Extraction (TCE) for Hybrid Cubes,” J. Eng. Appl. Sci., vol. 12, no. 8, pp. 2164–2169, 2017.

[25] M. F. Mushtaq, S. Jamel, K. M. Mohamad, S. K. A. Khalid, and M. M. Deris, “Key Generation Technique based on Triangular Coordinate Extraction for Hybrid Cubes,” J. Telecommun. Electron. Comput. Eng., vol. 9, no. 3–4, pp. 195–200, 2017.

[26] S. Jamel, M. M. Deris, I. Tri, R. Yanto, and T. Herawan, “HiSea : A Non Binary Toy Cipher,” J. Comput., vol. 3, no. 6, pp. 20–27, 2011.

[27] L. Granboulan, E. Levieil, and G. Piret, “Pseudorandom Permutation Families over Abelian Groups,” Fast Softw. Encryption, vol. 4047, pp. 57–77, 2006.

[28] A. Akhavan, A. Samsudin, and A. Akhshani, “A novel parallel hash function based on 3D chaotic map,” EURASIP J. Adv. Signal Process., vol. 2013, no. 1, pp. 1–12, 2013.

[29] J. Ahmad and F. Ahmed, “Efficiency analysis and security evaluation of image encryption schemes,” Int. J. Video Image Process. Netw. Secur., vol. 12, no. 4, pp. 18–31, 2012.

[30] A. Rukhin et al., “A statistical test suite for random and pseudorandom number generators for cryptographic applications,” Natl. Inst. Stand. Technol., pp. 1–82, 2010.

Web Service Testing Techniques: A Systematic Literature Review

Israr Ghani[1], Wan M.N. Wan-Kadir[2], Ahmad Mustafa[3]
School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia
Skudai, Johor Bahru, 81310, Malaysia

***Abstract*—These days continual demands on loosely coupled systems have web service gives basic necessities to deliver resolution that are adaptable and sufficient to be work at runtime for maintaining the high quality of the system. One of the basic techniques to evaluate the quality of such systems is through testing. Due to the rapid popularization of web service, which is progressing and continuously increasing, testing of web service has become a basic necessity to maintain high quality of web service. The testing of the performance of Web service based applications is attracting extensive attention. In order to evaluate the performance of web services, it is essential to evaluate the QoS (Quality of Service) attributes such as interoperability, reusability, auditability, maintainability, accuracy and performance to improve the quality of service. The purpose of this study is to introduce the systematic literature review of web services testing techniques to evaluate the QoS attributes to make the testing technique better. With the intention of better testing quality in web services, this systematic literature review intends to evaluate what QoS parameters are necessary to provide better quality assurance. The focus of systematic literature is also to make sure that quality of testing can be encouraged for the present and future. Consequently, the main attention and motivation of the study is to provide an overview of recent research efforts of web service testing techniques from the research community. Each testing technique in web services has identified apparent standards, benefits, and restrictions. This systemic literature review provides a different testing resolution to industry to decide which testing technique is the most efficient and effective with the testing assignment agenda with available resources. As for the significance, it can be said that web service testing technique are still broadly open for improvements.**

Keywords*—Quality assurance; web service testing; web service testing techniques; web service component testing

I. Introduction

Web services are growing in popularity in the software industry [1]. In order to achieve higher success rates several testing models, frameworks, tools, and techniques have been proposed and widely used where testing plays an important role to maintain the high quality of web services. However, it remains a challenging task to test web services, whether the type is SOAP (Simple Object Access Protocol) or REST (Representational State Transfer). Since there is a wide range of challenges, a comprehensive systemic literature review is presented to observe what the critical challenges are, with a focus on quality attributes and existing solutions to resolve the testing issues in web services.

Web services and micro services are software system. Both web services and micro services are utilized for coordination of various web applications using specific measures. Web services have advanced to the point that minimal effort is required for conveying information that is isolated between platforms and applications running on different functional systems and languages. Web services can be classified into two types: web service utilized in the intranet and the Internet. Web service given by the intranet is restricted to inside the concerned organization and is not accessible to the general public, while web services over the internet are available for all to utilize.

The benefit of intranet web service is that it is safe as the organization has full control. One benefit is that the service is only available to internal clients. With an internet web service, security [2-4] performance [3, 5, 6] efficiency [7, 8] reliability, interoperability [9] and accuracy [8, 10] need more concentration in order to obtain successful test rating. Mentioned parameters have become leading issues in web service testing due to its distinctive scalability [11] and confidentiality concerns.

Furthermore, real challenging task [12-14] in web service testing domain is that it has no Graphic User Interface (GUI) [15]. Thus, web service has no GUI to be verified and tested. Therefore, testing analysers involve some development capabilities and explicit tools [9, 16], framework [17-19], model [6, 11, 17, 20] or calculations to test GUI in web services. To put it plainly, testing of web services is not a straightforward job to be completed and require extraordinary capabilities, skills and innovation. Many tools, frameworks, approaches [20] and models have been introduced to improve web service testing. As a result, a few methodologies [10, 21, 22] are recommended to examine the nature of these techniques and tools. This research depends on essential analysis of two testing approaches: black box [23, 24] and white box [24] to test web service [25]. Current studies [8, 26-29] proposed techniques [8, 28, 29] and approaches for web service testing and produced an assessment to execute in order to obtain expected results.

For that reason, this study highlights those parameters which are not viably taken care of in these testing techniques. The results of the comparison between the primary studies will indicate [30] increasingly proficient testing strategies.

With the expectation of better testing quality in web service, this study presents an overview of testing techniques to assess which QoS parameters are important to give better quality affirmation. The focus of this study is to ensure that nature of testing can be encouraged from the present and future perspectives. Therefore, the primary consideration and inspiration of the study is to explore the related test techniques and provide an overview of the recent research in web service testing.

II. Related Work

AL Lemos and G Fraser [31, 32] also introduce the testing techniques to improve the quality of web service testing. This study introduces several testing approaches, models and tools to improve the web service quality better according to the industry requirements. In our primary study, we introduce 20 most used and trending testing techniques of web services where some authors suggested models and framework [7, 9, 11]. R. Casado and Z. U. Singhera [10, 33] present tools and frameworks against quality attributes such as performance interoperability, maintainability, reusability, efficiency and reliability. Jamshidi and Pooyan Pahl, Claus [34]explained different techniques and framework for SOA based service testing and mentioned how the service quality can be improved further.

III. Research Method and Motivation

In order to conduct a comprehensive analysis on testing of web service, we have defined research questions in Table I after analysis of selected primary study from Table I to Table IV stepwise.

Different models, approaches, frameworks and tools to recommend a resolution for web service testing have been described by most studies. These models and frameworks and tools have additional overhead and are not sufficient on multiple and different platforms while testing web service[34]. It has been observed that for multi dimension platforms, there are essential requirements for web service testing teams to work in this way [35]. In order to incorporate industry level web service testing, a comprehensive resolution[36] should be recommended to get the expected results.

The bottleneck in web service testing is the maturity of the exclusive parameters like reusability, performance, security, interoperability, maintenance, reliability and efficiency [37, 38]. Non-functional requirement (NFR) in web service testing [33] is neglected in previous literature reviews because of the nature of web service testing process [39]. There is a need for more empirical research in this area.

TABLE. I. Research Questions

Research Question (RQ)	Question Statement
RQ1	What are the issues in testing web services?
RQ2	What are the quality related aspects and concerns with respect to testing of web services?
RQ3	What are the existing solutions to solve the web services testing issues?
RQ4	How were existing testing techniques applied?

Some authors and studies have discussed automation of web service testing as an advantage of web service testing; some studies [40] suggest different automation testing techniques using tools and frameworks for web service testing. There is a lack of comprehensive research in the study for automation in web service testing. There is a need for more research and study to perform some extra investigation with an automation type of testing in web service. Based on the most common issues [41] and quality aspects according to the primary study, study and experiments of testing should focus on interoperability, performance, reliability and efficiency regarding the control of different testing techniques.

Existing tools and approaches [42, 43] are not sufficient to reveal structural errors to test web service [44]. This has motivated us to go through the literature and analyse the web service testing with respect to SOAP and RESTful and has led to the following research questions.

A. *Research Questions*

The primary interest of researcher is to identify the scope of future research activities for testing of web services. This systematic review identified the research study to make it more prominent for suggested study. Based on selected primary studies, the following research questions has been raised to make the web service testing techniques more effective and useful for the industry as well as for future researchers.

Following are the inclusion and exclusion criteria driving the research questions generated after selection of primary study.

B. *Inclusion Study Standards*

The criteria for inclusion of study are:

- Papers published from 2010 to 2019.
- Search regarding web service testing SOAP and RESTful.
- Search for papers that address specific testing problems in Document style XML and RPC.
- Search for papers that offer practical support for web service SOAP and RESTful testing technique.
- Papers that have quality related aspects and concerns with respect to web service quality assurance.

Studies selected in order to solve the web service testing issues.

C. *Exclusion Study Criteria*

The following steps are taken to exclude papers from primary study:

- Studies published before 2010.
- Research papers that not relevant to web service quality assurance.
- Title based research material that is problematic to classify is based on title search and referring to next step.

- After reading the abstract, the articles are identified as duplicate.

D. Study Selection Approach

In the selection of survey procedure, web service testing based work has been selected. Initially, digital databases are explored i.e. " 'Science Direct', 'ACM Library', 'IEEE Xplore', 'ISI Web of Science', 'Springer Link', 'Scopus' and 'Wiley' ".

In order to get desired results, manual search has also been executed in different sources corresponding to items of the scientific databases. To find the appropriate research studies that are not available in main database, Google Scholar was primarily used as a Meta - engine. Search for selected study was restricted to papers published between 2010 and 2019. As a result of initial queries, we obtained 674 published papers regarding web service testing techniques from 2010 to 2019.

Most of the sources contain important journals, symposiums, and workshop and conference proceedings within the web service domain. Search terms are mentioned in Table II. The queries were executed on different characteristics depending on the available options in database.

In primary study selection criteria, we make sure that published papers that were associated to web service testing techniques were included in search goal.

The search study is being extended instead of imperfect using the option (OR) web service AND testing, meaning that we submitted this search string in seven scientific databases as shown in Table III.

According to the provided information, the selected study was limited to the title, abstract, keyword to ensure that the results of the search string are appropriate in the context of the study.

Most of the sources include important journals, symposiums, and workshop and conference proceedings within the web service domain. Search terms are mentioned in Table II. The queries were executed on different characteristics depending on the available options in database.

In primary study selection criteria, we make sure that published papers that were associated to web service testing techniques were included in search goal.

The search study is being prolonged instead of imperfect using the option (OR) web service AND testing, meaning that we submitted this search string in seven scientific databases which available in Table III.

According to the provided enough information and after limited the selected study to the title, abstract, keyword to make sure, whether the results of the search string are appropriate in the context of the study or not.

By performing the search string in segment 1 of IEEE, a total 674 studies are found between 2010 and 2019. In segment 2, the study was reduced to 191 when title based execution criteria were applied. In segment 3, it is limited to 79 after eliminating the duplicate studies especially after reading the abstract. In segment 4, a total of 20 primary studies were selected after revising full text. Fig. 1 explains the procedure of selecting primary study criteria. The identifying primary studies are available in Table IV. The selection of primary study and review protocol has been well explained in this segment.

Fig. 1 presents the results of steps which provide details of the selection of selected primary study. After exploring the studies, 20 primary studies were selected spanning from 2010 to 2019 from conference, journals and book chapters.

Effective primary studies and 20 potential have been selected by relating the automatic and manual search method. These selected studies are associated directly with the question and interrogations related to answering the questions posed by our study. Selected study introduced technical papers, including SLR's which explored the role and testing approaches in web service in the past few years.

TABLE. II. DATABASE ON-LINE SOURCES

Library source	URL to access
Science Direct	http://www.sciencedirect.com
ACM Library	http://www.acm.org
IEEE Xplore	http://www.ieeexplore.ieee.org
ISI Web of Science	http://www.webofknowledge.com
Springer Link	http://www.springerlink.com
Scopus	http://www.scopus.com
Wiley	http://www.onlinelibrary.wiley.com

TABLE. III. SEARCH STRINGS, DATA SOURCES AND SELECTED STUDIES

Name	Results	Query String
ACM Library	34	Issues OR Challenges in web service testing
IEEE Xplore	94	Issues in web service testing OR Quality assurance OR attributes OR issues in web service AND challenges in web service testing
Web of Science	147	Issues in web service testing OR Quality assurance OR attributes OR issues in web service AND challenges in web service testing
Springer Link	29	Testing issues in web service OR quality assurance OR attributes OR issues AND testing of Web service
Science Direct	117	Testing Issues in Web Service AND micro service OR Challenges in Web Service Testing Issues OR API Testing Issues
Wiley	139	Web Service Testing Issues OR Challenges in micro service Testing
Scopus	114	Issues in Web Service Testing OR API Service Testing Issues
TOTAL	674	

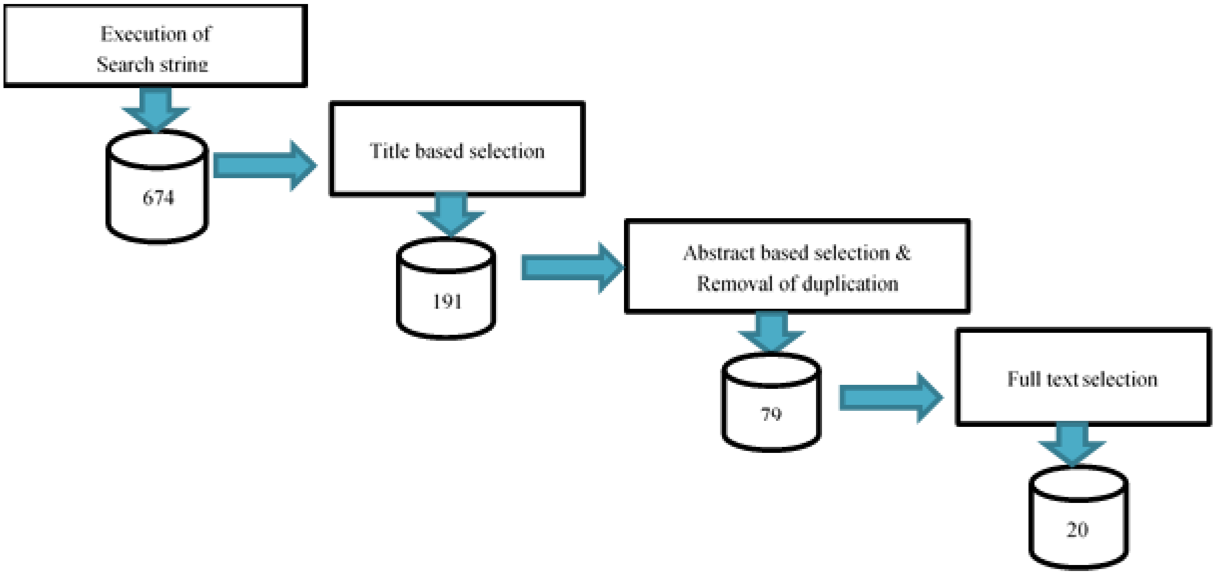

Execution of Search string
674
Title based selection
191
Abstract based selection & Removal of duplication
79
Full text selection
20

Fig. 1. Phase of Collection Method.

TABLE. IV. THE SELECTED PRIMARY STUDIES

Study ID	References	Description
S1	[38]	Specifications for Web Services testing
S2	[6]	Performance inquisition of web services using soap UI and JMeter
S3	[13]	Web Service testing Challenges and Approaches
S4	[22]	Testing Approaches for SOA and Web Service
S5	[29]	Semantic Web Services testing
S6	[46]	Regression Test Selection for Large-Scale Web Service Testing
S7	[47]	Functional testing of semantic web services
S8	[48]	Testing web service
S9	[17]	A Model-Based Framework for Cloud API Testing
S10	[20]	Black-Box Model-Based Regression Testing of Fail-Safe Behaviour in Web Applications
S11	[9]	An extensible tool for interoperability testing of web services
S12	[42]	A Reusable Automated Acceptance Testing Architecture for micro services
S13	[49]	Whitening SOA testing via event exposure
S14	[10]	A framework to test advanced web services transactions
S15	[50]	Vulnerability testing tools for web services
S16	[3]	Performance Analysis of Web Service Composition
S17	[5]	Performance Inquisition of Web Services
S18	[8]	An Abstract Transaction Model for Testing the Web Services Transactions
S19	[51]	Web Services Composition Testing Based on Extended Finite State Machine and UML Model
S20	[52]	Gremlin: Systematic Resilience Testing of Microservices

In Table IV is shown primary studies papers published from 2010 to 2019, which provides an outline of state of the art and categorization of some testing techniques.

The other challenge was to sort from selected testing approaches those related to associated study and which are appropriate to designing primary study conditions. In order to follow the above four segments, 20 most recently published and miscellaneous research papers are listed in Table IV. Furthermore, based on our quality parameter questions from selected primary studied and research exploration method, an analysis of each one has been performed individually.

E. Synthesis and Data Extraction

In order to extract appropriate material from selected primary study data, extraction have been performed to explain research questions on which research is based. To get the expected outcome, recommendations for such studies are to explore the introduction and methodology. Following is the required separated material from all selected primary studies:

- The source of selected primary studies, i.e. from conference, the title or journal.
- The structure of each study conducted.
- Method pros and cons of web services of both SOAP and RESTful testing technique.
- Supportive facts of web service testing approaches.
- Provision of tools for a web service testing technique.
- Necessary adoptive facts for web service testing techniques.
- The facts required to respond to the selected primary study questions.

The summary of data extractions is shown in Result and Discussion segment to explain their importance in terms of appropriate research. Furthermore, the explanation to classify the selected primary and other studies were incorporated and the categorization and execution process and its flow are well described as shown in Fig. 2. Narrative synthesis was used to elaborate the research methodology performed to get selected primary studies for analysis in Table IV.

This section explains the review protocol and procedure selected for the primary study. To evaluate and understand the specific area or study of interest, literature review is an efficient way to find the questions of respective research [53]. This procedure validates that the current research is reasonable. Consistent with recommendation mentioned by Kitchenham in Fig. 2, the literature review has three significant segments, which are planning, conducting and documenting. Requirement of the study is already mentioned in Section 1, and the description of research questions are also available in Table I. Data extraction procedures are explained in Fig 1. The outcome of selected primary study is given in Fig. 1 as well.

All research questions are appropriate to test and to quality the web service that is being explored through current approaches. Additionally, this study includes some parameters that are most common and important from selected primary studies for quality assurance.

In Fig. 2 is shown the description of a categorization pattern and extraction of data, including all phases of the producer which has been executed for selected primary study.

The classification of the selected primary study by year is shown in Fig. 3. The most published papers that have been produced and the abstract web service testing techniques research study were in 2018.

Fig. 3 presents the classifications of the major contribution by year. It shows that most papers were published for web services testing approaches in 2018. Similarly, Fig. 3 explains that there is a big discrepancy between the published papers from 2010 to 2019 regarding web service testing techniques or methodologies and tool supported techniques.

This study presents the analysis of the following parameters to answer all four research questions having one goal, which is how to achieve and progress the quality of web services testing for both SOAP and RESTful testing. Selected parameters that come through from primary study using search strings include security, performance, interoperability, maintainability, reusability, reliability and efficiency. The actual goal is also to make a thorough evaluation of most common issues while testing these parameters found in a selected primary.

To achieve these goals, according to RQ1, RQ2 and RQ3, we introduced a sequence of research questions as shown in Table I. We specified four research questions to analyse the primary study. The most common issues were raised while performed testing for selected parameters using existing techniques against web service testing and its type SOAP and RESTful and an explanation of these common issues and parameters is presented in Fig. 4.

RQ1 and RQ2 further elaborate the need for selected common issues with quality related aspects and concerns with respect to testing of web services. Parameters are divided intotwo categories, SLR’s and Journal, symposium, chapters. Most of the research of primary study includes these classifications according to common parameters using existing techniques, models, framework and approaches.

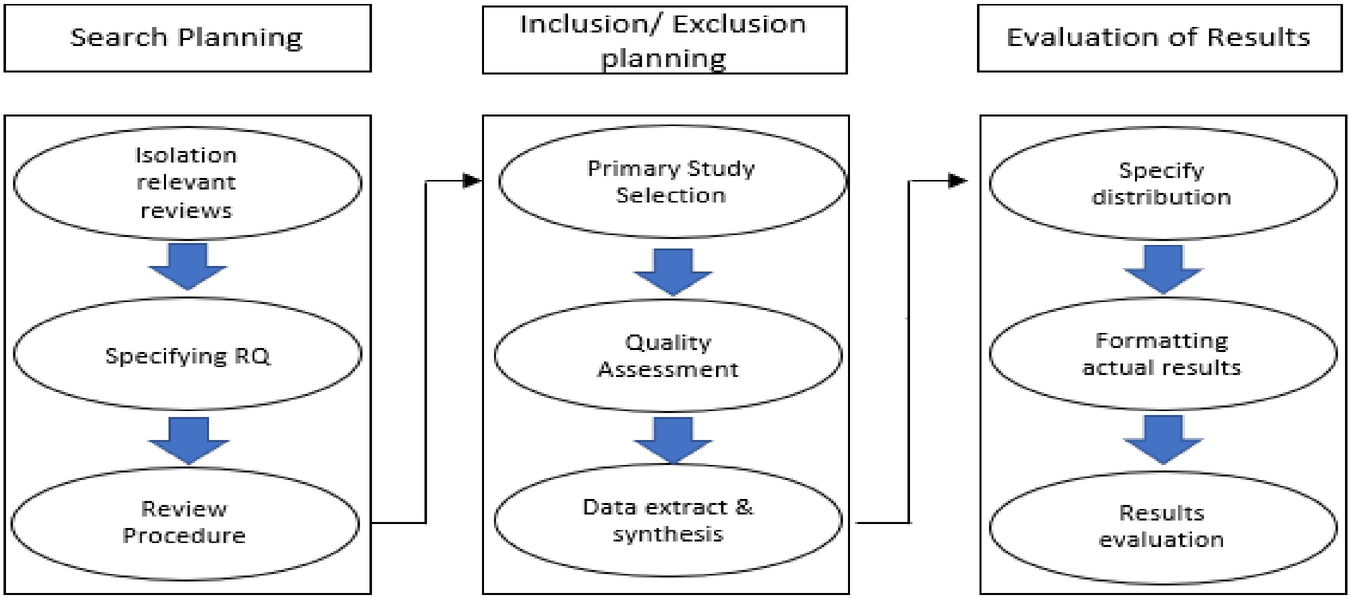


Fig. 2. Research Methodology from [45].

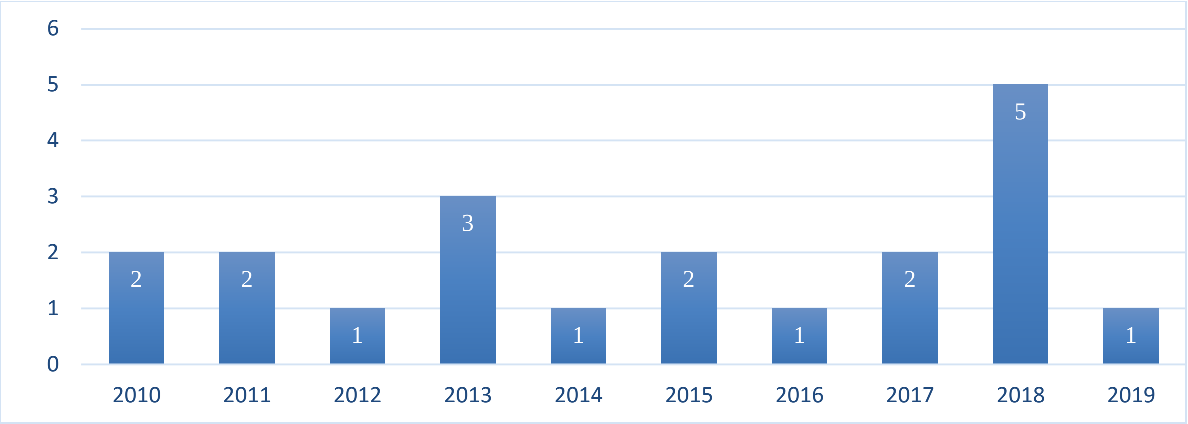


Fig. 3. Primary Study Year Wise Trend.

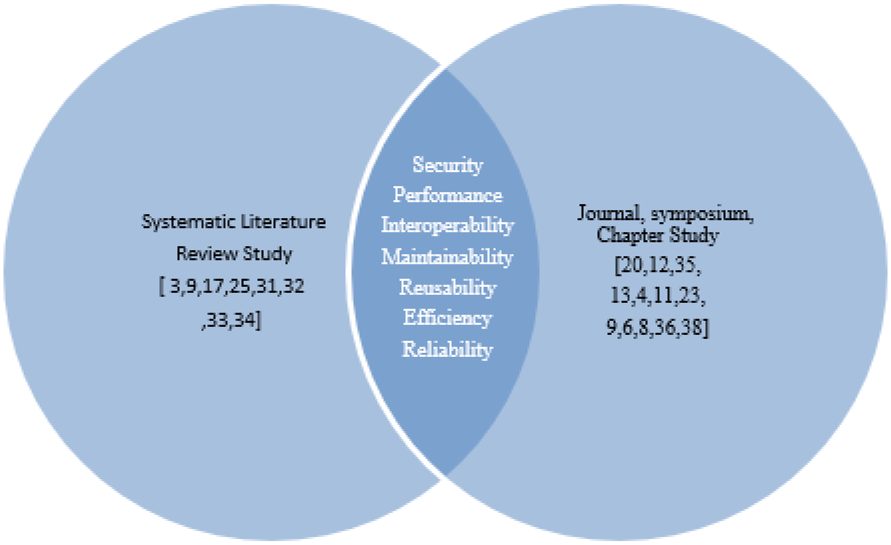

Systematic Literature Review Study [3,9,17,25,31,32 ,33,34]
Security
Performance
Interoperability
Maintainability
Reusability
Efficiency
Reliability
Journal, symposium, Chapter Study [20,12,35, 13,4,11,23, 9,6,8,36,38]

Fig. 4. Web Service Testing Issues in Common Parameters.

1) *RQ1"What are the issues web service testing?"*

The following explanation provides the details of each quality attribute which the most common issues are found in selected primary study from Table IV.

a) Security is a basic term in software and web service testing in the context of software or web service testing to play out its errand by using a base measure of resources while successfully communicates either within the system or other systems. Most of the time, in order to exchange data and information, a system must interact with other systems using different mechanism. For web service security testing [54, 55], this is a critical element, as the mechanism or tool should have the ability to make sure the security interrelate and then manage with remaining components.

b) Performance specifies the ability of the mechanism to ensure performance quality and also is a measure of how fast a request of service can be completed. Performance also ensures testing [5] the sensitivity under the specific workload and stress situations. There are two important types of performance testing which are stress and load testing. Stress testing is usually conducted to test the performance of a system under extreme capacity, increasing its maximum load. Load testing is performed by executing the system under explicit predefined load capacity to realise its performance and efficiency.

c) Interoperability refers to the ability of the system to interrelate efficiently with different frameworks whenever needed [9]. It also refers to making sure while testing that web services and system are interoperable to implement services between the different development environments. Usually a framework is needed to perform data trade through different components to associate with additional frameworks. This is a vital element, as the instrument should have the capacity to collaborate and arrange with outer parts.

d) Maintainability This attribute assesses if the code of the system can be retained to address weaknesses, alter or develop in the upcoming changes in the system. Code practicality sense can be sophisticated by following high coding standard, doing proper documentation of code and keeping up consistency to the extent programming dialect and coding benchmarks [56]. From future perspective code modularization and Object-Oriented Programming can likewise help in code.

e) Reusability This term specifies the ability of the testing mechanism to accomplish reusability of web software and web service testing. Under explicit embedded the same functionality and circumstances, in this study reusability deals with sensitivity of the web service testing [42]. Multiple fluctuating constraints can be reused from the earlier version's test cases of web service. So, testing of reusability can reduce work effort to provide solution with constraint standards for those variables while testing the latest changes in the system and application.

f) Efficiency This attribute shows the capacity of web service and software system to play out and run by using a base measure of assets. Efficiency in software and web service testing will assess if the web service testing makes the most efficient consumption of framework assets like processor parameter, RAM, transfer speed and so forth. An instrument consuming excessive plate space or memory shows a low proficiency [57]. Thus, the testing model, framework or technique which is being used must sure prompt results while performing efficiently as a testing technique that does not respond in time to produce the expected results can be considered as inefficient.

g) Reliability this parameter represents the ability of the application to execute the given assignment by consuming a lowest volume of assets, which signifies the capacity of a web service testing to execute the given necessary tasks under mentioned circumstances within the respective time period. The purpose of the reliability [58] is to make sure the overall measures of web service testing and its quality are maintained. Reliability will evaluate the usage of system resources like RAM, processor capacity, bandwidth while testing web service in this study.

The mentioned testing features and terms are the direct and valuable testing opinions, and must thus be valued and realised for web service quality testing.

2) RQ2 What are the quality related aspects and concerns with respect to testing of web services?

Fig. 5 is concerns the analysis of selected studies to demonstrate the RQ2 quality related aspects, which are security, performance, interoperability, maintainability, reusability, efficiency, reliability with respect to RQ1 as shown in Fig. 4. When dealing with web services testing, additional significance is given to applied research leading to resolutions than to theoretical exploration like literature reviews. Due to the large number of web service testing approaches, research studies that relate to the experiments applied in the lab and are not yet used in an actual industry environment. Based on implemented experiments, it can be concluded that this study has many open exploration problems that require further investigation.

There seems to be an overflow in web service testing to address common quality related aspects as shown in Fig. 4. This is because multiple testing techniques that might be reasonably comparable to the comprehensive level are stated with different titles. This could be a result of freshness of study testing area because of deficiency of recognised terms. Identifying approaches that follow related techniques are not credible with particular information mentioned in the abstract as well as introduction and conclusion level. This is one of the prominent explanations that inspire us to extend this study to a systematic literature review.

3) RQ3 What are the existing solutions to solve the web service testing issues?

It is useful for researchers to recognise latest solution capacities that have not as yet been explored comprehensively. This requires a focus on the limitations, disadvantages and gaps of the current testing techniques. However, the answer to RQ3 well be explained in following Result and Discussion section.

The outcome in testing of web service is providing a comprehensive overview of the current state of the research from this mapping study. The purpose of this study is to produce a thorough taxonomy of web services testing techniques in the context of selected parameters from primary study.

Existing testing technique and proposed solution in this study is done through different methodologies which are frameworks, models and tools available in Table VII. Other testing strategies, which are regression testing and load testing, security testing has been obtained from these existing techniques. The RQ3 question is useful for the testers to select testing techniques before providing service. To answer this question, our attention will be on parameters and testing techniques utilized in each study.

4) RQ4 How were existing testing techniques applied?

Test generation process and techniques focus on the procedure which served to produce the web service auto test proposed to ensure the quality of test selection methods. By applying the manual test, auto test, graph search algorithms, model-checking, reasoning and other manual test is the process to ensure web service testing. This is done to place greater emphasis on the major contribution for each selected analysis mentioned.

According to the earlier mentioned parameters, there is a need to explain how existing approaches and techniques are applied to meet the requirements of RQ4. In this review authors suggested some models, tools and techniques to produce product to end user. Therefore [43, 59] introduce new approaches for testing web service. Author in [25] presents a black box testing technique to improve the quality. A research conducted by [3] introduced performance testing approach, which tests and assesses the general behaviour or efficiency of the structure of the system under business hours and several solutions have been proposed in primary study by the authors.

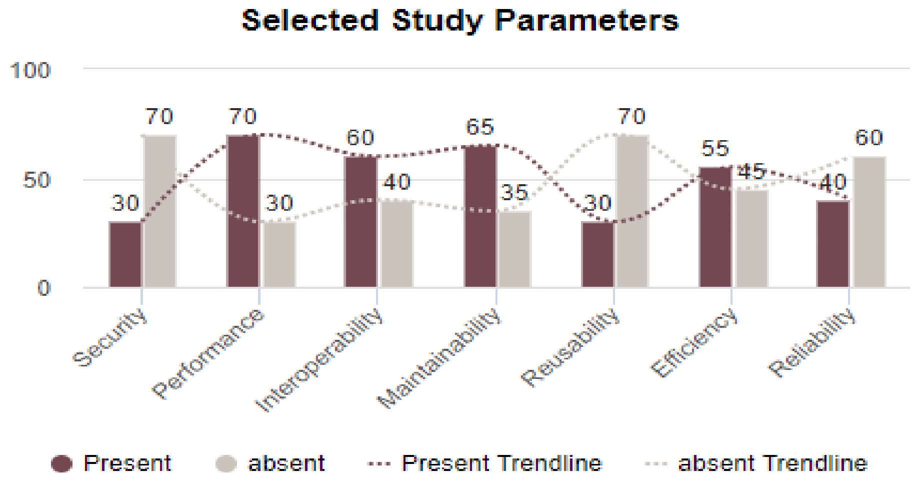


Fig. 5. Results and Findings of Evaluating Studies.

IV. Result and Discussion

A. Web Service Quality and Analysis

To formulate the results to address research questions for web service testing techniques with a solution, Quality parameter outcomes have been introduced in Table V. The signs "✓" and "x” have been introduced for present and absent, respectively, to display parameters that the explicit approaches represent.

For web services quality assurance, the testing mechanism for security, performance, interoperability, maintainability, reusability, efficiency, reliability have been analysed in detail.

This segment discusses the results of the study mentioned in Table V. Collectively, each subcategory attempts to provide appropriate answers to the questions we have defined in Table I. It also explains the implementation of web services testing and how these techniques are introduced by using different methods to implement these approaches, model and frameworks.

From Table V it is known that almost 70% of testing approaches claim to be appropriate for security, performance, reusability and more than 60% interoperability, maintainability and reliability and reusability efficiency of web service testing. It is significant for the web service to work in an orderly manner for appropriate functioning of web services and its testing as it has the competency of effective incorporation and procedure for performance, interoperability, maintainability, reusability and reliability. These parameters show the value of being effective and some of the technique provides solutions for efficiency. Most of the studies do not discuss efficiency in their experiments using different testing techniques as shown in Fig. 5. More than 55% of testing approaches are focused to ensure the efficiency of testing in web service.

This study proposed and presents the testing approaches as well as code maintainability, security, efficiency, reusability, reliability, interoperability, performance, which are important to ensure the quality of web service testing for both SOAP and RESTful. Proposed techniques consider interoperability, which is crucial for web service testing. Interoperability and security are vital elements to ensure the quality of the system during testing, but most studies do not consider and discuss it as per analysis in Fig. 5. The same is true for reliability and reusability while testing web service where even 40% techniques are not followed to ensure better quality testing. It is essential to make sure that all functional and non-functional desires are satisfied, whether these requirements are predefined are or not while web services testing.

The major contribution of results are shown Table VI, where the leading role to ensure the quality of web services testing is performed using maintaining security, interoperability, maintainability, reusability, efficiency, reliability parameters using different techniques, tools, framework or models. To test the mentioned parameters according to the selected primary studies, Table VI uses research methodology as shown in Fig. 2.

Table VI presents the major classifications in this area according to RQ3 and RQ4 resolutions using frameworks, tools and existing approaches as a result of growth in the web service testing production and fluctuation in the performance of business. The speed of testing procedure has increased according to the primary studies based on web service testing structures and influence listed in percentage in Fig. 6.

The classification of the study includes framework, model or tool support in Fig. 6. It can be seen that 70% of the works are supported by different techniques or methods, with 20% of the works proposing tools which developed for web service testing purpose and 40% of studies introducing different frameworks models suggested by the authors.

In order to find keywords and planning that improve the outcome, in this paper is selected conditions distinguished by reading the abstracts of the papers. The designated keywords and ideas are taken from related available research to present a categorization of structure that replicates an awareness of the base of the research study input. In this manner, we have classified the papers into categories (called Major Contribution) mentioned in categorization pattern in Fig. 6. This procedure was completed using spreadsheets for each facet and its scopes.

TABLE. V. SELECTED PRIMARY STUDIES FOR ANALYSIS

Parameters	**[29]**	**[39]**	**[4]**	**[6]**	**[7]**	**[9]**	**[10]**	**[11]**	**[14]**	**[18]**	**[21]**	**[22]**	**[34]**	**[35]**	**[36]**	**[37]**	**[32]**	**[40]**	**[41]**	**[38]**
Security	x	✓	✓	x	x	x	✓	x	✓	✓	x	x	✓	✓	x	✓	x	x	x	x
Performance	x	✓	✓	✓	✓	x	x	✓	✓	✓	x	x	x	x	x	✓	x	x	✓	✓
Interoperability	x	x	x	x	x	x	✓	x	✓	✓	x	x	✓	x	✓	✓	x	x	x	x
Maintainability	x	✓	x	x	✓	x	x	x	x	✓	x	x	x	x	x	✓	✓	x	x	✓
Reusability	x	x	x	✓	x	x	x	x	x	x	✓	x	x	x	x	x	✓	x	x	x
Efficiency	x	x	x	x	x	✓	x	✓	✓	x	✓	✓	x	✓	x	x	x	✓	✓	✓
Reliability	✓	x	x	x	x	✓	x	✓	✓	✓	✓	✓	x	x	✓	x	x	✓	✓	✓

TABLE. VI. MAJOR CONTRIBUTION IN THE LITERATURE

Contribution Types	**References**
Technique/Method	[4],[29],[6],[14],[22],[34],[35],[36],[37],[32],[40],[41],[38]
Tool	[39],[10],
Framework/Model	[21],[18],[40],[9],[7],[11]

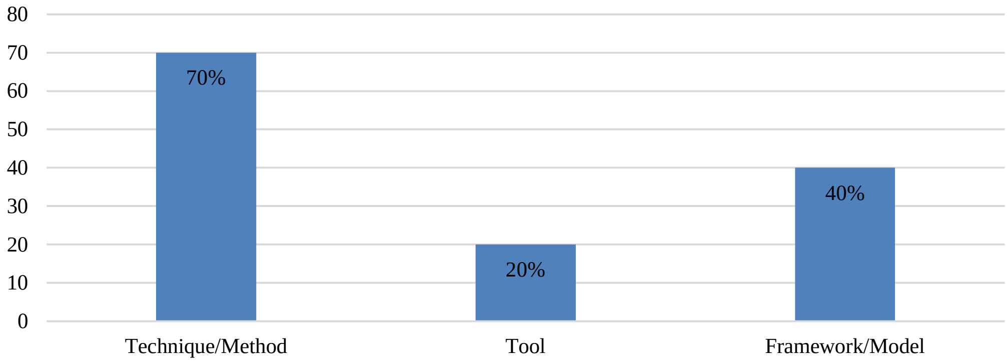


Fig. 6. Classification of the Major Contribution.

The classification of selected primary study that has SOAP and RESTful web service type support are presented in Fig. 7, which shows a sample illustration of the types of existing web service. As mentioned earlier, basically, there are two types of web services, SOAP and REST. However, SOAP is further classified into Remote Procedure Call (RPC) style and Document XML style.

It is shown that almost 70% of the studies supported SOAP while 30% explained for Restful. The authors use Remote Procedure Call (RPC) style and Document XML style because Document XML style is more effective, efficient and reusable while performing testing, specifically for interoperability, performance, security and maintainability shown in Table V using different techniques or methods presented in Fig. 8.

In order to use discussed prototypes, Fig. 8 shows the distribution of the major contribution over the web service types. Note that the analysis only contributes selected primary study available in Table IV. We can identify the dissimilarity between the quantity of SOAP and RESTful studies in supported studies as SOAP is makes it much easier to communicate between the interfaces and is easy for developers to deploy.

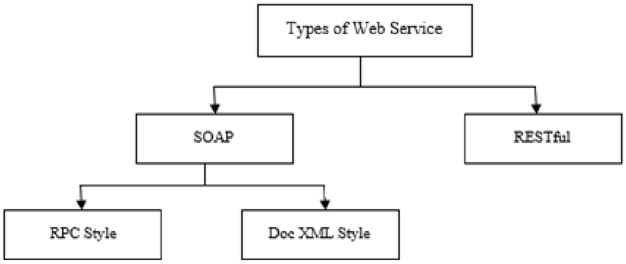


Fig. 7. Types of Web Service.

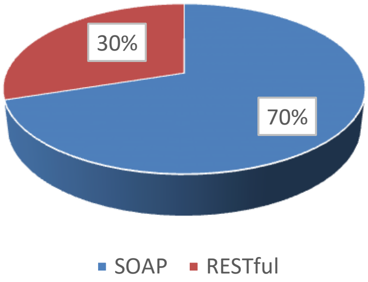


Fig. 8. Major Contribution of Web Service Types.

TABLE. VII. CONTRIBUTION OF WEB SERVICE TYPES FOR ANALYSIS

Web Service Types	[29]	[39]	[4]	[6]	[7]	[9]	[10]	[11]	[14]	[18]	[21]	[22]	[34]	[35]	[36]	[37]	[32]	[40]	[41]	[38]
SOAP	✓	✓	✓	✓	✓	x	✓	✓	✓	✓	✓	✓	✓	x	✓	✓	✓	✓	✓	x
Restful	x	x	✓	x	✓	✓	x	✓	x	x	✓	x	x	✓	✓	x	x	x	x	✓

TABLE. VIII. MAJOR CONTRIBUTION OF WEB SERVICE TYPES IN SELECTED PARAMETERS FOR ANALYSIS

Web Service Types	Security	Performance	Interoperability	Maintainability	Reusability	Efficiency	Reliability
SOAP	[29],[39],[4],[14], [18],[21] [35],[37]	[39],[4],[6],[14], [21],[34],[35], [37]	[29],[39],[10], [18],[36], [37]	[39],[10],[11], [21],[34],	[29],[35],[35], [32],[41], [41]	[11],[37], [40],[41],[38]	[29],[6],[9], [14],[34],[35] [37],[40],[38]
RESTful	[4],[18],[21], [35]	[4],[18],[21]	[18],[36]	[11],[21],	[32]	[11],[38]	[9],[32],[38]

By using primary study available in Table IV we have performed analysis against each study of web service type that has been formulated in tabular way as presented in Table VII. The signs "✓" and "x" have been introduced to indicate the presence and absent of discussed primary study. Here is the major contribution of web service types in Table VII. It presents the main influence for web service types, SOAP and RESTful against selected parameters according to the selected primary studies.

The major contributions are shown in Table VII, which presents the main influence in web service testing for SOAP and RESTful done through maintaining security, interoperability, maintainability, reusability, efficiency, reliability issues using different techniques, tools, framework or models to test the mentioned parameters according to the selected primary studies.

For consistency, the current study explicitly selected parameters presented in Table VIII, which reports what web service type is used against which parameter. The explicit selected parameters presented in Table VIII address the web service types.

B. Challenges of Web Service Testing

According to the analysis of earlier primary studies, it has been observed that measures to ensure the quality of service must be satisfied according to the RQ1 and RQ2. Additionally, some other system architectural parameters to improve the quality of web service testing are also significant, for example security and efficiency measures. Most of the quality assurance parameters have been discussed in Table VIII to improve testing of web services using different techniques.

There are possible that the one testing approach will be the best in one testing standard and the worst in testing another. But in order to reduce the gap between testers and developer, it is important to improve the quality of the product because ultimately only client is going to use the service irrespective of the product after testing. The other issue, which is a challenging one, is increasing the quantity of APIs, as there are different services against each API [50]. Moreover, continuous maintenance of web service is also problematic to ensure the quality as it is possible that every user may not maintain or upgrade his service. Therefore, it is important to improve and introduce such techniques to ensure better quality better that can work for all kinds of services rather than maintaining quality for the different versions individually. The most significant points are to assure performance, security, interoperability, maintainability, reusability, reliability and efficiency of web services as well as for micro service through testing.

RQ3 is focused to assure better quality in web services. In addition to above, there should be more extensive work required to:

- Introduce the quantity of testing efforts in terms of frequency that should be required to make the quality better.
- Introduce the necessary levels of testing consistent with each perspective (i.e. Provider, Integrator, Developer, and Third Party as well as for End User).
- Possible testing techniques at Integrator and third-party level.
- Few quality features are mentioned in selected studies to improve quality performance for every type of web service or micro services testing which are helpful in finding testing area.

Twenty primary studies are shown in Table IV and we classified them into the four categories of web service testing on the basis of latest techniques, The primary study we have selected includes the use and discussion of black box testing technique in (13 papers), regression testing selection (2 papers), performance testing explained in (4 papers), and white box testing discussed in (1 paper) only, which are discussed in detail in review method section.

After performing analysis and evaluation to conclude the current state of the art of useful (black box, white box, regression testing and load testing) in a web service in order to isolate the challenges related with these existing techniques, tools and frameworks which introduced in Table IX.

Fig. 9 shows the classification of testing technique and testing level facets according to Table IX. Therefore, it has been observed in the different search studies that 50% of the primary study discusses and concentrates on black box testing. Similarly, many studies proposed regression testing regarding test case concerns. Only one study is about white box testing and some authors discussed load testing in selected studies as shown in Fig. 9. These selection criteria are selected according to RQ4 from selected primary studies.

Fig. 9 shows the results of testing level used in primary studies and it seems that 50% of the selected studies propose black box testing while 35% propose regression testing, 10% study's authors suggested load testing and 5% for white box testing in order to ensure the quality of web service testing of security, performance, interoperability, maintainability, reusability, efficiency and reliability. Therefore, the purpose of using these suggested testing techniques from black box methodology can be concluded as:

- Easy to use: It is easy to use as the testing team is not required to understand the internal structure of the system. Therefore, test case generation process is a relatively stress-free work.
- More rapid testing: In this testing, the testers must only interact with system's Graphic User Interface (GUI), no time is exploited in analyzing the structure of the code. This kind of practice saves the lot of time of testers and makes the testing process much easier and faster.
- Simpler process: When the testers must deal with large and composite systems, some techniques or methods like black box testing make the testing process simpler because testers are only concerned with the valid and invalid inputs and the expected output.

The following studies are associated with the relevant SOAP [51], [52] and RESTful testing [50], [20], [53] techniques. The summary of the study stated that explanations of web service testing can be further improved using existing techniques, models, frameworks and tools of testing by supporting the responsibilities that are required for the capability and understanding of a service tester. However, most of the existing web service testing techniques, frameworks, models and tools [46], [10] still need improvement as most suggested approaches are in their early stages. Most of the suggested techniques only provide the results of initial conclusions, which indicates that there is a need for more concentration and need to discuss further improvements in web service testing area, specifically on quality parameters like security [3], performance [4], interoperability [10], maintainability [46], reusability [32], efficiency [38], and reliability [11].

The primary study is focused on functional and non-functional web service requirements. It also concluded that it is possible to derive testing from specifications of selected parameters and the challenges of web service testing are, in general, the same as in the testing of traditional web services. Furthermore, it is recommended that instructions for future exploration of the study focus on existing techniques, tools and frameworks to improve web service testing. Transformation techniques relative to research the different works should be initiated to explore new techniques [15, 52, 59] to identify web service testing techniques. The testing of web service with the selection of the most important parameters assured its quality whenever the specification of web service changed to develop or introduce different techniques, tools, framework, and models to test quality parameters. Still, the majority were not properly validated or evaluated during testing of web service.

C. *Test Method used for Testing Web Service*

This review found a number of known testing techniques completed for web service such as regression testing, model based testing, load testing, rule based testing and performance testing. Built-In Operational Testing, Passive Testing, different test case methods, tools and models introduced in primary study make the web service testing better and improve its quality. [60] Study introduces structural and operational testing technique, which is obviously a white box testing system as it associated with structure of web services programming. [61] Presents regression testing techniques to ensure the latest program or code alteration has not affected current structures.

In [62] authors have presented technique for reporting of test results which is an XML- based technique. As per our investigation, most of the testing approaches are state or model based which are operational for testing of web services.

For web services, there is a requirement of testing from different interpretations with less information on user end. There should be a testing technique which can be applicable for every sized service as the services can be small and large. Therefore, different testing techniques, tools and models are more appropriate to use because of time effective technique and ease of use. The following figure presents the same idea to cater selected parameters using existing testing techniques from selected primary study.

Fig. 10 shows the combination between selected study parameters, testing techniques and testing level facets. However, we have to ensure good quality testing techniques which have been introduced in recent years and also included primary selected studies. Below is the comparison to ensure that selected study tools or framework offer good quality testing techniques.

TABLE. IX. MAJOR CONTRIBUTION OF TESTING TECHNIQUES IN WEB SERVICE

Testing Types/levels	**References**
Black Box	[29],[21],[22],[34],[35],[36],[37],[18],[10],[32],[11] [39],[9]
Regression Testing	[29],[14], [40]
Performance Testing	[7],[4],[6],[41],
White Box	[38]

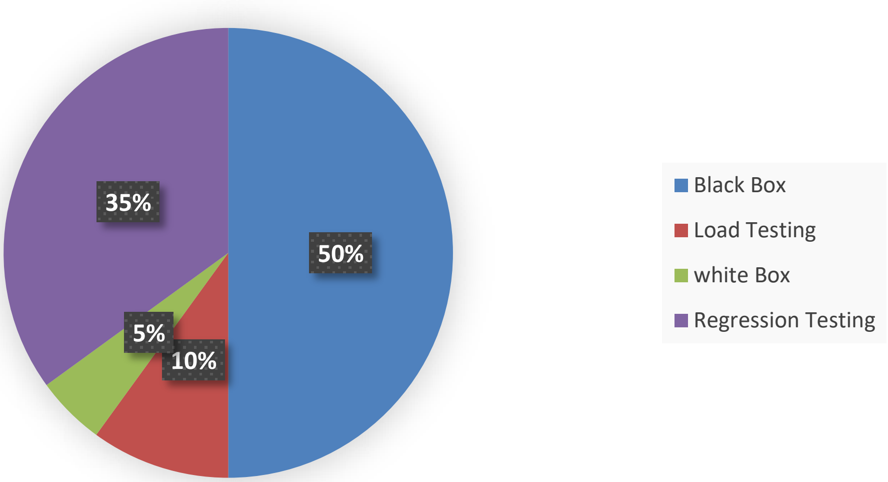


Fig. 9. Contribution of Solution Techniques with Respect to Existing Studies

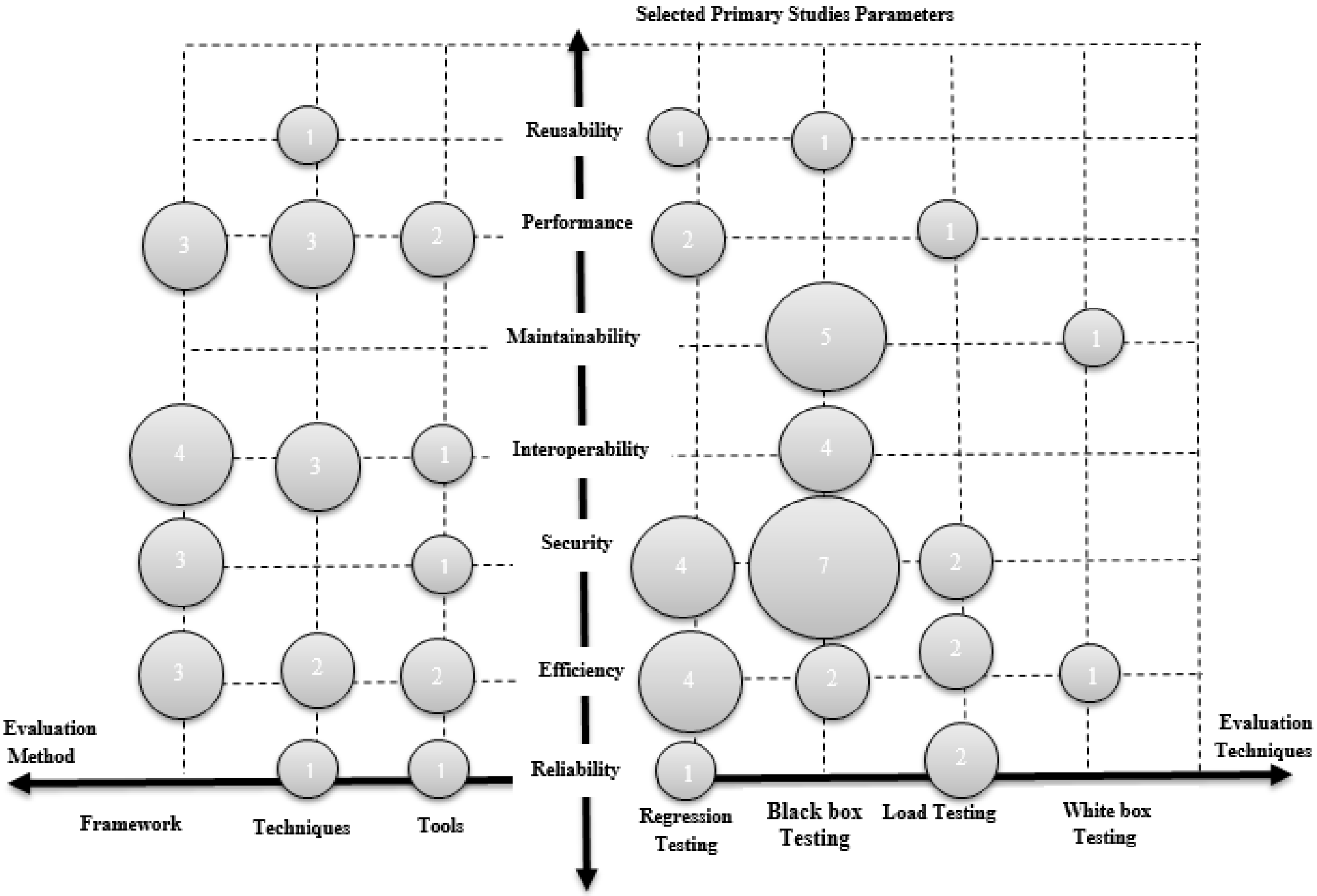


Fig. 10. The Relationship between Primary Study, Testing Techniques and Testing.

The symbols "✓" and "x" are used, respectively, for presence or absence from the selected primary study. The major contribution of web service quality testing attributes are shown in Table X, which indicates the main influence for quality studies of web service testing techniques according to the selected primary studies. Table X shows quality attributes that we have defined in our study for good and valuable quality testing techniques. When we discuss quality attributes of web service testing, we are basically analysing what has been claimed in the primary study, the study method used and how the research question was answered using defined quality parameters.

TABLE. X. MAJOR COMPARISON OF QUALITY STUDIES

Selected Primary Study	Comprehensive literature review	Proposed Techniques	Conducted Experiment	Presented Results	Compared Results with previous study	Claimed to Improve Results	Contributed to knowledge
[38]	x	✓	x	✓	✓	x	x
[50]	✓	✓	✓	✓	✓	x	✓
[3]	x	x	✓	✓	x	x	✓
[5]	x	✓	✓	✓	✓	x	x
[6]	x	✓	✓	✓	✓	x	✓
[8]	x	✓	✓	x	x	✓	✓
[9]	✓	✓	✓	✓	x	x	x
[10]	x	✓	✓	✓	x	x	✓
[13]	x	x	✓	✓	x	x	x
[17]	✓	x	x	x	x	x	✓
[21]	✓	✓	✓	✓	x	✓	✓
[63]	x	✓	✓	✓	x	x	x
[29]	x	✓	✓	✓	x	✓	✓
[62]	✓	x	✓	✓	x	x	✓
[47]	x	x	x	x	x	x	✓
[48]	✓	✓	✓	✓	✓	✓	✓
[42]	x	✓	✓	✓	x	x	✓
[51]	x	x	✓	✓	x	x	x
[52]	✓	✓	✓	✓	x	x	✓
[49]	✓	✓	✓	✓	x	✓	✓

As shown in Table X, there were many limitations and weaknesses in selected primary studies as no comparison was made with previous studies while experiments, Presented Results and Claimed to improve the results. Also, most of the works concerned with quality attributes of quality testing and prioritization and evaluation are related to selected primary study and RQ2 and RQ3.

As shown in the individual analysis, most of the works deal with the existing techniques using defined quality attributes to fulfil the requirements of RQ3 against quality parameters.

Therefore, during analysis we found that only 8 primary studies are comprehensive literature where 14 studies are proposing approaches or techniques, while 17 studies conducted experiments while testing web services and all 17 presented results. However, these results were not compared with previous studies which shows that every researcher continued with their experiments without contributing knowledge and only 5 studies compared the results with previous studies. Results with previous studies need more concentration in order to contribute knowledge to the study to improve the final results, where 5 studies claimed to improve approaches and techniques. 14 primary studies contributed knowledge and suggest some appropriate solution frameworks mentioned in Table X. It will help the testing industry to use existing techniques and approaches, models and frameworks. Fig. 11 shows the trend of the major contribution of quality studies which is self-explanatory according to Table X.

It is observed in Fig. 11 that there are some research trends such as comprehensive literature review, presented results, compared with previous studies and claimed to improve the results which were not satisfactory until now. Most of the articles recommended resolutions for current study, but still there is a need of evaluation study. Additionally, there is a need of assessment and standard research results in proposal. Further organizational research is needed to explore the real-time process in web service testing.

D. Selected Quality Study Attributes

Following is the detailed explanation of each quality attribute to make the web service testing better. There might be some other quality attributes for analysis, comprehensive literature review. According to our selected primary study, we are considering only quality attributes as shown in Table XI. To understand which study provides better quality for testing web service, the intention of the best solution for testers is to make the service better for the end user.

The reasons for using different suggested methods, techniques, frameworks or tools under the defined quality attributes-based method has been defined in Table X are as follows because web services are mixture of interfaces.

E. Comparison of Research Findings

Below is the comparison of previous studies and current study in comparison of research findings. A total eight literature reviews were compared with current study. According to the research parameters, results are expressed to address research questions for web service testing techniques with a solution. Parameters already stated in results of primary studies from 20 actual research paper's results are mentioned in Table V in tabular form. The symbol "✓" and "**x**" are specifically introduced for presence and absence, respectively, to present that the explicit studies address the constraints and previous studies to compare with current study.

Table XI showed the difference the comparison between previous studies literature reviews and SLR. It is shown that previous SLR's did not introduce all parameters, where this study selected common issues while testing or issues not discussed in previous studies. Some studies are providing details of experiments and techniques as a solution only against interoperability or reliability. Other studies provide solutions for security and performance only using existing techniques. It is need to cater to all existing quality parameters to make the quality of web service testing better.

From the results shown in Table XI, we have suggested following implications to make the web service quality better.

- The researcher will obtain awareness of different Web Service testing techniques in the testing environment.
- The solutions (testing, integrations and other testing events) so far established for web service testing. These resolutions use different frameworks, model, tools that will help testers in adopting while testing Performance, Interoperability, Maintainability, Reusability, Reliability, and Efficiency.
- Use of frameworks and tools increase the simplicity while proceedings in web service testing.
- The progressive and destructive effect of web service testing will support experts and organizations to implement suggested testing frameworks, models and tools.
- In web service testing, tools and frameworks support play an important role.
- Different testing techniques, models, frameworks play an important role in Web Service testing.
- The systematic literature review supports the testers to select frameworks, tools and models in testing conditions to establish web service testing in an easy and secure way.

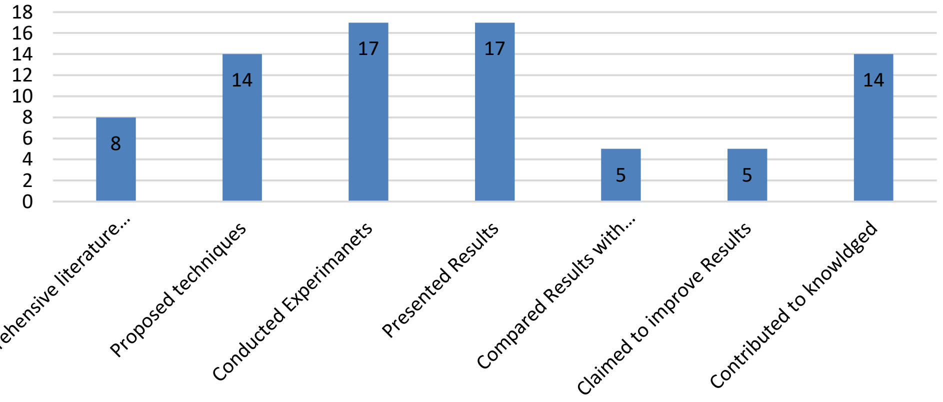


Fig. 11. Major Contribution of Quality Study.

TABLE. XI. COMPARISON BETWEEN PREVIOUS STUDIES AND CURRENT STUDY

Overall Studies	Security	Performance	Interoperability	Maintainability	Reusability	Efficiency	Reliability
[38]	x	x	✓	x	x	x	✓
[29]	x	✓	X	x	x	x	✓
[6]	x	✓	✓	x	x	✓	x
[62]	✓	x	X	✓	x	✓	✓
[47]	x	x	✓	x	x	✓	✓
[37**]**	x	✓	X	x	x	x	✓
[13]	✓	✓	✓	x	x	✓	✓
[48]	✓	✓	✓	x	x	x	✓
Current Study	✓	✓	✓	✓	✓	✓	✓

V. Threat to the Validity in Research

The conclusion of systematic literature review research is consistent but there is a possibility of validity and bias threats for the duration of selection of primary study, conduct analysis or result phase of the SLR study [45]. Step of collection of primary study or exploration approach recognises maximum possible studies available in the literature by avoiding bias. There is possibility to skip some researches because of grey areas of study associated with literature such as contribution of solution techniques and relationship between primary study, testing techniques and testing levels. The research work is associated with multiple software testing communities, quality assurance, information system and simple object access (SOA) and web service testing. The method to identify the primary study is completed by considering internal validity, bias and external validity proposed and recommended by Cochrane Reviewers Handbook [42]. For further validation, the study is established on the implementation of an approach to promote the review. The transparency of the study will assist the researchers to evaluate the consistency of outcomes accurately.

The outcome of systematic literature review demonstrates that validation of the research is required in all kinds of support as shown in Table X. Assessment required some practical case studies and skilled reports in this capacity, which encourage researcher and testers.

VI. Concluding Remarks and Future Perspective

Web service testing is a significant issue that must be considered sensibly; testing must be broad, comprehensive and automated to all significant levels (unit, Black box, white box, regression, load, component, and system level testing). In order to evaluate the performance of web services, it is essential to evaluate the QoS (Quality of Service) attributes such as interoperability, reusability, auditability, maintainability, accuracy and performance to improve the quality of service. Different approaches, frameworks, models and tools among the main influences towards seven parameters, which are security, performance, interoperability, maintainability, reusability, reliability, and efficiency to evaluate better quality of web service testing. According to our primary study, 70% of web service testing is performed based on different testing techniques instead of frameworks or models to ensure better quality of web service testing. In the utilization of web services regarding different functional and non-functional requirements of the system, such framework and models provide better ways. Evaluation and testing at three main levels, such as end to end, service to service and interface to interface testing is important to ensure better quality of web services because of their extraordinary structure. This is because the most concerned aspect is the user who requires a better graphic user interface (GUI) with new change and provide compatibility, interoperability and performance with user machine. Further significant aspects that require attention may include traceability, compatibility, and complexity. Integrity and effectiveness should also be included to ensure better quality of web service. Furthermore, this study can be improved further by adapting Model based approaches; Semantic based approaches for existing and remaining quality parameters, increasing the quantity of selected primary studies by covering all the digital databases. Additionally, this study can be extended by including some extra testing techniques based on the following parameters: traceability, compatibility, complexity, scalability, auditability, effectiveness, integrity and inconsistencies in web service testing.

Acknowledgment

The authors would like to express their deepest gratitude to Universiti Teknologi Malaysia (UTM) and Ministry of Education Malaysia (MOE) for their financial support under the Research University Grant Scheme (Vot number Q.J130000.2516.19H64).

References

[1] Kumar, A. K. Pandey, and M. Singh, "A novel testing framework for SOA based services," International Conference for Convergence of Technology, I2CT, Pune, India, 2014.

[2] M. I. P. Salas and E. Martins, "A Black-Box Approach to Detect Vulnerabilities in Web Services Using Penetration Testing," IEEE Latin America Transactions, no. 3, 2015.

[3] S. Jing and D. Yu-Yue, "Performance analysis of web service composition based on stochastic well-formed workflow," The International Conference on Networked Computing, Proceeding, Gyeongju, South Korea, 2010.

[4] R. Selvam and A. Senthilkumar, "Webservice based vulnerability testing framework," Proceeding of the IEEE International Conference on Green Computing, Communication and Electrical Engineering, ICGCCEE, Coimbatore, India 2014.

[5] S. Radhakrishna and M. Nachamai, "Performance inquisition of web services using soap UI and JMeter," 2017 IEEE International Conference on Current Trends in Advanced Computing, ICCTAC, Bangalore, India, 2018.

[6] S. Shridevi and G. Raju, "A literature survey on the performance evaluation model of semantics enabled web services," 2018.

[7] A. A. Nacer, K. Bessai, S. Youcef, and C. Godart, "A Multi-criteria Based Approach for Web Service Selection Using Quality of Service (QoS)," Proceedings IEEE International Conference on Services Computing, New York, USA, 2015.

[8] R. Casado, J. Tuya, and M. Younas, "An abstract transaction model for testing the web services transactions," Proceedings IEEE 9th International Conference on Web Services, ICWS, Washington, DC, USA, 2011.

[9] I. A. Elia, N. Laranjeiro, and M. Vieira, "ITWS: An extensible tool for interoperability testing of web services," Proceedings IEEE International Conference on Web Services, ICWS, Anchorage, AK, USA, 2014.

[10] R. Casado, J. Tuya, and M. Younas, "A framework to test advanced web services transactions," 4th IEEE International Conference on Software Testing, Verification, and Validation, ICST, Berlin, Germany, 2011.

[11] M. Silic, G. Delac, I. Krka, and S. Srbljic, "Scalable and accurate prediction of availability of atomic web services USA," IEEE Transactions on Services Computing, 2014.

[12] Z. Mansor and E. Edwin, "Issues , Challenges and Best Practices of Software Testing Activity," Proceedings of the 14th International Conference on Applications of Computer Engineering (ACE '15),Seoul, South Korea, 2015.

[13] S. Azzam, M. N. Al-kabi, and I. Alsmadi, "Werb Service Testing Challenges and Approaches," 2012.

[14] A. Mendoza and G. Gu, "Mobile Application Web API Reconnaissance: Web-to-Mobile Inconsistencies & Vulnerabilities," 2018.

[15] S. P. Ma, C. Y. Fan, Y. Chuang, W. T. Lee, S. J. Lee, and N. L. Hsueh, "Using Service Dependency Graph to Analyze and Test Microservices," International Computer Software and Applications Conference, Japan, 2018.

[16] C. F. Liao, C. J. Cheng, K. Chen, C. H. Lai, T. Chiu, and C. Wu-Lee, "Toward A Service Platform for Developing Smart Contracts on Blockchain in BDD and TDD Styles," IEEE 10th International Conference on Service-Oriented Computing and Applications, SOCA, Kanazawa, Japan, 2017.

[17] J. Wang, X. Bai, L. Li, Z. Ji, and H. Ma, "A Model-Based Framework for Cloud API Testing," 2017.

[18] N. Pande, A. Somani, S. Prasad Samal, and V. Kakkirala, "Enhanced Web Application and Browsing Performance through Service-Worker Infusion Framework," IEEE International Conference on Web Services (ICWS), San Francisco, CA, USA, 2018.

[19] X. Chen, Z. Ji, Y. Fan, and Y. Zhan, "Restful API Architecture Based on Laravel Framework," Journal of Physics: Conference Series, 2017.

[20] A. Andrews, A. Alhaddad, and S. Boukhris, "Black-Box Model-Based Regression Testing of Fail-Safe Behavior in Web Applications," Journal of Systems and Software, 2018.

[21] Z. Chen, L. Shen, and F. Li, "Exploiting Web service geographical neighborhood for collaborative QoS prediction," Future Generation Computer Systems, 2017.

[22] F. Besson, P. Moura, F. Kon, and D. Milojicic, "Bringing Test-Driven Development to web service choreographies," Journal of Systems and Software, 2015.

[23] S. Mehta, G. Raj, and D. Singh, "Penetration Testing as a Test Phase in Web Service Testing a Black Box Pen Testing Approach," in Smart Computing and Informatics: Springer, 2018.

[24] M. Khanna, N. Chauhan, and D. Sharma, "A Novel Approach for Regression Testing of Web Applications," International Journal of Intelligent Systems and Applications, 2018.

[25] K. Incki, I. Ari, and H. Sözer, "A survey of software testing in the cloud," in IEEE Sixth International Conference on Software Security and Reliability Companion, Gaithersburg USA, 2012.

[26] A. Souri, A. M. Rahmani, and N. Jafari Navimipour, "Formal verification approaches in the web service composition: a comprehensive analysis of the current challenges for future research," International Journal of Communication Systems,, 2018.

[27] J. Hu, X. Chen, Y. Cao, and L. Zhu, "A Comprehensive Web Service Selection Algorithm on Just-in-Time Scheduling," 2016.

[28] A. Askarunisa, K. A. J. Punitha, and A. M. Abirami, "Black Box Test Case Prioritization Techniques for Semantic Based Composite Web Services Using," International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, Tamil Nadu, India, 2011.

[29] J. B. De Souza Neto, A. M. Moreira, and M. A. Musicante, "Semantic Web Services testing: A Systematic Mapping study," Computer Science Review, 2018.

[30] M. Ficco and M. Rak, "Stealthy denial of service strategy in cloud computing," IEEE Transactions on Cloud Computing, 2015.

[31] A. L. Lemos, F. Daniel, and B. Benatallah, "Web service composition: a survey of techniques and tools," ACM Computing Surveys (CSUR), 2016.

[32] S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Cortés, "A survey on metamorphic testing," IEEE Transactions on software engineering, 2016.

[33] Z. U. Singhera and A. A. Shah, "Extended web services framework to meet non-functional requirements," Workshop proceedings of the sixth international conference on Web engineering - ICWE Tokyo, Japan, Japan, 2006.

[34] P. Jamshidi, C. Pahl, N. C. Mendonça, J. Lewis, and S. Tilkov, "Microservices: The journey so far and challenges ahead," IEEE Software, 2018.

[35] H. Rusli, S. Ibrahim, and M. Puteh, "Testing Web Services Composition: A Mapping Study," Communications of the IBIMA, 2011.

[36] D. Tosi and S. Morasca, "Supporting the semi-automatic semantic annotation of web services: A systematic literature review," Information and Software Technology, 2015.

[37] P. Moretto, P. Rossaro, S. A. MacArthur, and D. Truffelli, "Systems and methods to identify and classify performance bottlenecks in cloud based applications," ed: Google Patents, 2018.

[38] E. I. Nabil, "Specifications for Web Services Testing: A Systematic Review," World Congress on Services, New York, NY, USA, 2015.

[39] W. A. Cidral, T. Oliveira, M. Di Felice, and M. Aparicio, "E-learning success determinants: Brazilian empirical study," Computers & Education, 2018.

[40] N. E. Ioini and A. Sillitti, "Open Web Services Testing," World Congress on Services, Washington, DC, USA, 2011.

[41] F. J. Barroso, J. H. Cass, M. R. Deckert, M. J. Saylor, and A. Skwersky, "Automatic detection of problems in a large-scale multi-record update system and method," ed: Google Patents, 2019.

[42] M. Rahman and J. Gao, "A reusable automated acceptance testing architecture for microservices in behavior-driven development," 9th IEEE International Symposium on Service-Oriented System Engineering, IEEE SOSE, San Francisco Bay, CA, USA, 2015.

[43] T. Zhang and Q. Yao, "An Approach of End User Regression Testing for Semantic Web Services," International Conference on Management and Service Science, 2011.

[44] I. Baldini et al., "Serverless computing: Current trends and open problems," in Research Advances in Cloud Computing: Springer, 2017.

[45] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature reviews in Software Engineering," Engineering, 2007.

[46] H. Zhong, L. Zhang, and S. Khurshid, "TestSage: Regression Test Selection for Large-Scale Web Service Testing," in 12th Conference on Software Testing, Validation and Verification (ICST) Xi'an, China, China, 2019: IEEE.

[47] A. Tahir, D. Tosi, and S. Morasca, "A systematic review on the functional testing of semantic web services," The Journal of Systems & Software, 2013.

[48] M. Bozkurt, M. Harman, and Y. Hassoun, "Testing Web Services : A Survey."

[49] C. Ye and H. A. Jacobsen, "Whitening SOA testing via event exposure," Transactions on Software Engineering, 2013.

[50] N. Antunes and M. Vieira, "Designing vulnerability testing tools for web services: approach, components, and tools," International Journal of Information Security, 2017.

[51] C.-s. Wu, "The Web Services Composition Testing Based on Extended Finite State Machine and UML Model," Fifth International Conference on Service Science and Innovation, Kaohsiung, Taiwan, 2013.

[52] V. Heorhiadi, S. Rajagopalan, H. Jamjoom, M. K. Reiter, and V. Sekar, "Gremlin: Systematic Resilience Testing of Microservices," Proceedings - International Conference on Distributed Computing Systems, Nara, Japan, 2016.

[53] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," 2007.

[54] A. Mendoza and G. Gu, "Mobile Application Web API Reconnaissance: Web-to-Mobile Inconsistencies &amp; Vulnerabilities," 2018.

[55] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," Journal of network and computer applications, vol. 34, no. 1, pp. 1-11, 2011.

[56] C. William, D. Poshyvanyk, M. K. Moran, C. Eduardo, and B. Cardenas, "Development , Testing and Maintenance of Android Apps : Challenges , Tools , and Future Directions," IEEE/ACM 5th International Conference on Mobile Software Engineering and Systems (MOBILESoft), Gothenburg, Sweden, 2018.

[57] M. A. Jamil, M. Arif, N. Sham, A. Abubakar, and A. Ahmad, "Software Testing Techniques: A Literature Review," 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M) Jakarta, Indonesia, no. November, 2016.

[58] W. Ha, "Reliability Prediction for Web Service Composition," pp. 7-10, 2017.

[59] M. J. Kargar and A. Hanifizade, "Automation of regression test in microservice architecture," 4th International Conference on Web Research, ICWR , Tehran, Iran, 2018.

[60] M. M. Eler, M. E. Delamaro, J. C. Maldonado, and P. C. Masiero, "Built-in structural testing of web services," Brazilian Symposium on Software Engineering ,Salvador, Bahia, Brazil, 2010.

[61] T. Masood, A. Nadeem, and S. Ali, "An Automated Approach to Regression Testing of Web Services based on WSDL Operation Changes," IEEE 9th International Conference on Emerging Technologies (ICET), Islamabad, Pakistan, 2013.

[62] S. Sim, "Formal Testing Approaches for Service-Oriented Architectures and Web Services : a Systematic Review ∗ e Takeshi Endo."

Deep Transfer Learning Application for Automated Ischemic Classification in Posterior Fossa CT Images

Anis Azwani Muhd Suberi[1], Wan Nurshazwani Wan Zakaria[*, 2], Razali Tomari[3]
Ain Nazari[4], Mohd Norzali Hj Mohd[5], Nik Farhan Nik Fuad[6]
Faculty of Electrical and Electronic Engineering (FKEE) [1, 2, 3, 4, 5]
Universiti Tun Hussein Onn Malaysia (UTHM)
86400 Parit Raja, Batu Pahat, Johor, Malaysia
UKM Medical Centre, Jalan Yaacob Latif[6]
Bandar Tun Razak, 56000 Cheras
Kuala Lumpur, Malaysia[6]

***Abstract*—Computed Tomography (CT) imaging is one of the conventional tools used to diagnose ischemic in Posterior Fossa (PF). Radiologist commonly diagnoses ischemic in PF through CT imaging manually. However, such a procedure could be strenuous and time consuming for large scale images, depending on the expertise and ischemic visibility. With the rapid development of computer technology, automatic image classification based on Machine Learning (ML) is widely been developed as a second opinion to the ischemic diagnosis. The practical performance of ML is challenged by the emergence of deep learning applications in healthcare. In this study, we evaluate the performance of deep transfer learning models of Convolutional Neural Network (CNN); VGG-16, GoogleNet and ResNet-50 to classify the normal and abnormal (ischemic) brain CT images of PF. This is the first study that intensively studies the application of deep transfer learning for automated ischemic classification in the posterior part of brain CT images. The experimental results show that ResNet-50 is capable to achieve the highest accuracy performance in comparison to other proposed models. Overall, this automatic classification provides a convenient and time-saving tool for improving medical diagnosis.**

***Keywords*—Deep learning; ischemic stroke; posterior fossa; classification; convolutional neural network; computed tomography; medical diagnosis**

I. Introduction

Ischemic stroke is a condition where there is a blood clot in the blood vessel which causes a blockage to the area of the brain [1]–[3]. Recent studies have shown that ischemic stroke is the third leading causes of mortality across the world, accounting for about 74.8% of the patient are experiencing their first episode of stroke in general [4]. The mortality rate of ischemic is also increasing with 3 million deaths globally [5]. There have been tons of ischemic are missed at the early stage of diagnosis which requires an accurate diagnosis to defeat the mortality rate. The delay of early diagnosis and treatment would stimulate the growth of acute ischemic which then reducing the chances rate of the patient to survive [6]. To date, Computed Tomography (CT) is the most applicable tool for rapid screening of ischemic in medical emergency [3][6][7]. Despite that, the diagnosis of ischemic in CT slices of Posterior Fossa (PF) remains as the most challenging area, which demands the expertise of radiologist [8]–[10]. The diagnosis process can be monotonous and time-consuming. Generally, the human brain consists of PF which covers two major components; Cerebellum and brainstem. The cerebellum is the primary site for body coordination and movement while the brainstem has been implicated for respiration. The beam hardening artifacts produced by thick bone and inadequate contrast resolution usually limits the performance of CT in PF slices [9]. False positive cases usually occur because of the behavioural nature of acute ischemic in loss of gray-white matter differentiation which nearly similar to normal tissues. The sensitivity of CT in the first 24 hours of ischemic inspection in PF slices achieves a low performance with only 41.8% [10]. Fig. 1 shows the example of ischemic case in PF region.

Computer Aided Diagnosis (CAD) has been widely developed in the medical image analysis to provide support for the radiologist in the decision-making process [11]–[14]. There are various techniques applied in the CAD to classify normal and ischemic CT slices using traditional Machine Learning (ML) based method. Typically, these CAD systems are starting with the pre-processing, followed by hand-crafted feature extraction and finally the classification steps to discriminate normal and abnormal slices [11][14][15]. Texture, intensity features and wavelet transform are the examples of features which have been extracted to be fed into Support Vector Machine (SVM), Artificial Neural Network (ANN) and K-Nearest Neighbour (KNN) as the input [11][14]–[16].

At present, Deep Learning (DL) has made a breakthrough in medical image classification owing to its superior performance in solving a wide range of problems [17]. Contrary to the traditional ML methods, DL is capable to directly learn the features of an image without a hand-crafted feature extraction step. The emergence of DL in medical image classification with diseases such as Alzheimer, brain tumour ischemic infarction and dementia are tremendously increasing with high performance [18]. However, only very recently have other researchers started to investigate the possibility of ischemic image classification with acute condition based on the DL method [12][19][20]. To the best of author knowledge, this is the first study to explore the challenge of classifying normal and ischemic primarily in PF slices using DL.

*Corresponding Authors

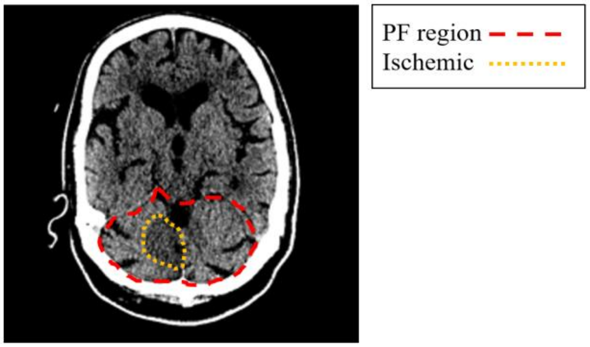


Fig. 1. Ischemic Stroke in PF.

Generally, the huge amount of CT image data affect the computational cost to train the DL network [21]. Convolutional Neural Network (CNN) is the most popular DL architecture [21][22]. Through the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), various pre-trained models of CNN have been developed with high boosting performance [23]–[25]. Pre-trained models can be another way of overcoming the problems of training CNN with huge data since the Graphics Processing Unit (GPU) hardware demands a high-cost. This is because the previous pre-trained models are able to apply the knowledge of transfer learning to another domain. Those models are only required to train the last dense layer since they use the similar weight of convolutional layers from the previous pre-trained.

There are several attempts which can be enumerated for deciding a patient with an acute ischemic stroke. These are usually intensive and lack of robustness due to the unknown stroke symptom onset [26]. Therefore, Pereira et al. [19] employ a CNN classifier for brain ischemic CT images by combining with Particle Swarm Optimization (PSO). They use CIFAR-10 and ImageNet architecture as a deep convolutional network model for classification. Chin et al. [20] develop a CNN network to classify ischemic using image patches. They use two convolutional layers, one max pooling layer and a single fully connected layer. More recently, Dourado et al. [12] develop a classifier for ischemic and haemorrhage stroke by combining ML-based classifier with CNN. They use the CNN as feature extractor instead of the hand-crafted feature extraction.

Despite all efforts, no specific study exist today that can provide a classification of normal and ischemic in PF slice. It is important to note that only 20% of ischemic stroke is represented by posterior [27]. This eventually influences the vast amount of training data needed for DL. Thus, this study employs and evaluates the pre-trained models as deep transfer learning in CNN towards classifying normal and ischemic in PF slices of CT. Early diagnosis of ischemic in PF is extremely important to prevent the brainstem infarction incidence which can impact the prolonged state of the patient [10]. The significant contributions of this study can be listed as follows to address this unaccounted phenomenon: (a) comparative study of pre-trained models to classify normal and ischemic mainly in PF slices of CT, and (b) presentation of cross validation to improve the performance of the predictive models. The following paper is organised as follows: (a) Section 2 describes the state-of-the-art of CNN models; Section 3 presents the research method used including brain dataset and research workflow; Section 4 shows the experimental results for initial training and *k*-fold cross validation, followed by Section 5 which summarises the findings.

II. Deep Transfer Learning Model of CNN

The pre-trained network is shown to operate effectively in the DL field. This is particularly important for small training dataset cases. The pre-trained models appear to possess relatively good performance with previous knowledge of large scale data of the existing model. This section describes a background on three popular pre-trained models (VGG-16, GoogleNet and ResNet-50) which have been proven to be excellent in medical image classification [28][29].

A. VGG-16

VGG-16 is the CNN model developed by Simonyan and Zisserman [25] with the best performance of 92.7% top-5 test accuracy in ImageNet. Fig. 2 illustrates the architecture model of VGG-16. This was developed and tested on a number of 1000 classes. The input required for the convolution layer is RGB with a fixed size of 224×224. The input image is passed via a stack of convolutional layers with ReLU activations in which filters are applied with very small receptive fields of 3×3. Meanwhile, the stride of convolution is fixed to 1. In the context of spatial pooling, five max-pooling layers are performed after the convolutional layers. Each of the max-pooling is typically used with a window of 2 x 2 pixels and a stride of 2. The stack of convolutional layers is followed by three fully connected (fc) layers and finally, the softmax layer. The ability of VGG to perform through an increase of effective network receptive field is one of the major advantages due to stacking stage of multiple convolution layers with small kernels as well as limiting the number of parameters [28]. For a number of years, this model in literature has been prevalent to the area of medical image classification such as mammograms [30].

B. GoogleNet

GoogLeNet is developed by Szegedy et al. [24] which delivers the best performance in ILSVRC 2014 with an error rate of 6.67%. The model is implemented using the strategic approach of stacking nine Inception modules. The implementation presents the advantage of employing optimal local sparse structure using dense components of a multi-layer network [28]. Particularly, the Inception module has included four branches. The first branch applies 1×1 convolution to perform a linear transformation to each input channel. Meanwhile, the second and third branches of GoogleNet exploit the principle of including 1×1 kernel convolution with kernel sizes of 3×3 and 5×5, respectively. The idea behind this is to perform a dimensionality reduction before applying max-pooling and convolution with 1×1 kernel in the fourth branch. Meanwhile, the purpose of using max-pooling is to reduce the feature maps dimensionality. The final stage of the architecture highlights the implementation of average pooling layer, fc layer, drop out and softmax loss as the classifier. The details of these structures are depicted in Fig. 3.

This study is supported by Universiti Tun Hussein Onn Malaysia and UKM Medical Centre under TIER 1 Vot H203 and GPPS Vot U954 Research Grants.

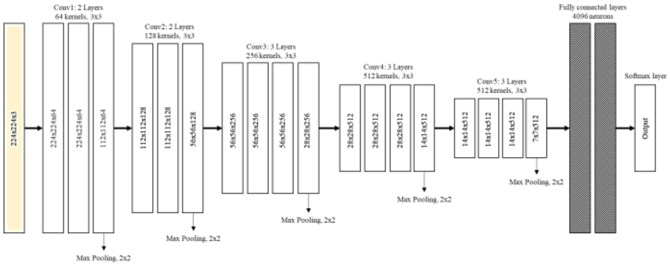


Fig. 2. VGG-16 Architecture [28].

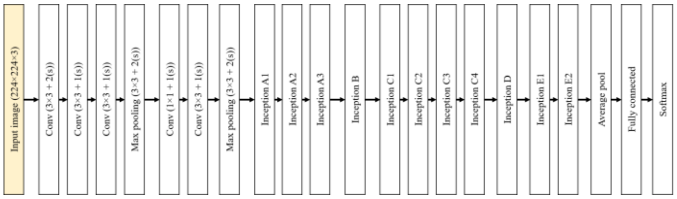


Fig. 3. GoogleNet Architecture [31].

C. ResNet-50

ResNet-50 or Residual Network is introduced by He et al. [23] in ILSVRC 2015. Similar with VGG-16 and GoogleNet, ResNet-50 also has been trained on the ImageNet database of 1000 categories with the superior performance of 3.57% error rate. ResNet-50 architecture is composed of convolutional layers, pooling layers and multiple residual layers with each consisting residual blocks of convolutional layers and batch normalization layers [30]. The configuration of ResNet-50 particularly consist of four residual layers, a dense layer and followed by the output layer with softmax activation function [30]. The design and implementation of each convolutional layer in a residual block are added with shortcut connection before propagating the output to the subsequent block as shown in Fig. 4. ResNet-50 provides an advantage in delivering a remarkable speed-up convergence as compared to other pre-trained models [18].

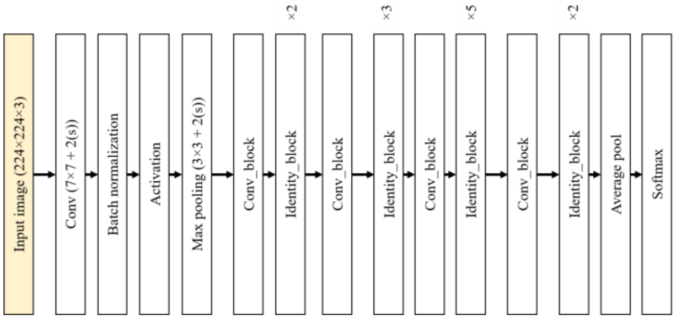


Fig. 4. ResNet-50 Architecture [31].

III. Research Method

As aforementioned, the main contribution of this work is to evaluate the performance of deep transfer learning under the context of ischemic classification in PF of the CT images. Fig. 5 illustrates the proposed workflow which the training and validation images undergo 1) Digital Imaging and Communications in Medicine (DICOM) image conversion, and 2) Data augmentation and 3) CNN pre-trained model. The deep transfer learning of CNN model has been implemented with MATLAB R2018B software. All experiments are performed on Intel® Core™ I7-7500U processor with 2.90GHz and 8GB RAM platform.

A. Brain Dataset

Brain CT images are acquired from the UKM Medical Centre. The images are scanned using Aquilion One by Toshiba scanner. The dataset contains normal and abnormal images. The abnormal images belong to acute ischemic in PF region. Only images with PF are used from the database except for non-PF slices. The size of all images is 512 × 512 pixels in an axial plane. 400 of images are used for training and validation. Among these images, 200 contains ischemic in PF and the rest of 200 are normal images. Fig. 6 shows a few sample CT images of the brain dataset. The ischemic is clinically confirmed and marked (green region) by an experienced radiologist.

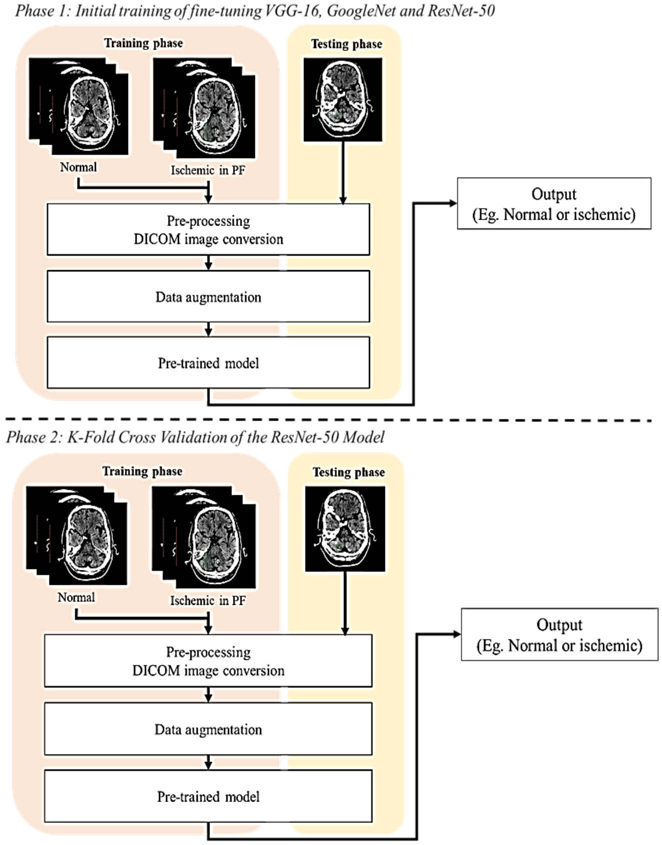


Fig. 5. The Overall Proposed Workflow.

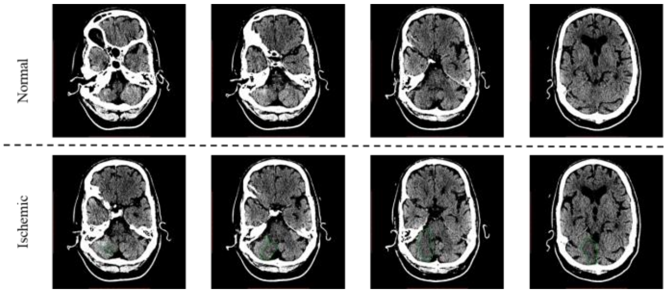


Fig. 6. Samples of Normal and Ischemic (Marked with Green Region) in PF of Brain CT Images.

B. Pre-Processing DICOM Image Conversion

The pre-processing operation starts with converting the 16-bit of DICOM image into 8-bit of grayscale range using Equation 1 and 2 for visualisation purpose. In order to enhance the visibility of the normal and abnormal tissues, the window setting of window width (W_w) and window center (W_c) must be in the range of +40 to +80 Hounsefield unit (HU) [32]. It is important to note that both W_w and W_c of 40 HU, recommended by the radiologist provides a solution that satisfies the visibility of acute ischemic in CT image [33].

$$W_{c(PV)} = \frac{W_{c(HU)} - R_i}{R_s} \quad (1)$$

$$I_{out} = (I_{max} - I_{min}) \times \frac{I_{in} - (W_{c(PV)} - \frac{W_w}{2})}{W_w} \quad (2)$$

where I_{out} –Output value of grayscale intensity (Output image), I_{in} –Input value of pixel (Input image), I_{max} – Maximum grayscale intensity value ($I_{max} = 255$), I_{min} – Minimum grayscale intensity value ($I_{min} = 0$), $W_{c(PV)}$ – Window center in PV, $W_{c(HU)}$ –Window center in HU, $W_{w(HU)}$–Window width in HU, PV–Pixel value, R_i –Rescale intercept, R_s–Rescale slope.

C. Data Augmentation

Training a deep learning network for classification demands a large training image. Image augmentation usually will be applied to boost deep learning performance by creating training images in different augmentation techniques [34]. Simultaneously, this technique also helps to prevent overfitting specifically for small dataset [18]. By having a small dataset, the network sometimes is unable to adapt to the new data. Therefore, in this stage, the augmentation technique which is horizontal flipping will be applied to increase the number of training images. Each image will create one (1) random artificial image as illustrated in Fig. 7.

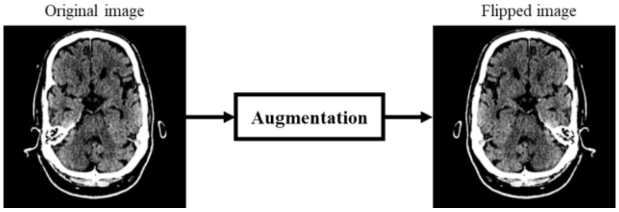


Fig. 7. Image before and after Augmentation Process.

D. Phase 1: Initial Training of Fine-Tuning VGG-16, GoogleNet and ResNet-50

The pre-trained models; Firstly, VGG-16, GoogleNet and ResNet-50 are trained and their respective performances are assessed based on the training and validation phase. The fine-tuning technique is applied in which the pre-trained models are trained and the connection weights are transferred to be adapted to the desired task. All layers with initial weights except fully connected layer are retrained without freezing. The dataset is divided into 80-20% for training and testing respectively. The CT images also are re-sampled to 224 × 224 pixels resolutions to match the input requirement of deep transfer learning models. Previous CNN pre-trained models are trained with input channels of red, green and blue. In this experiment, the input layer of 224×224×3 is replicated with a new layer of 224×224×1 since the image is in a grayscale form. Subsequently, the final layers of the pre-trained models are replaced with new layers. Fine tune the deeper layer is important for the models to learn the specific features based on the new dataset and classes. Throughout the training, Adam optimizer has been implemented to reduce loss function. The initial learning rate is set to 1e $^{-4}$ and the model is trained for 3 epochs with a batch size of 20 and Squared Gradient Decay Factor (SGDF) of 0.99. After the training, the best pre-trained model is selected based on performance metrics for the validation set.

E. Phase 2: K Fold Cross Validation of the ResNet-50 Model

In phase 1, ResNet-50 model has achieved the best performance. Further details of the results can be found in Section IV (A). The best predictive model will be properly assessed through 5-fold cross validation in phase 2 to verify that model biasing and generalization errors are counteracted. This type of validation will evaluate the trained model using each of the four partitions and the remaining one is under validation. Thus, the ResNet-50 model is trained five times and the average accuracy of 5-fold cross validation is computed.

IV. Experimental Results and Discussion

In this section, a detailed analysis of experimental setup results is discussed. The performance of predictive models have been evaluated through traditional metrics performance; precision, recall, F1-score, specificity, accuracy and processing speed. The traditional metrics to assess the performance of the proposed method are derived based on true positive (TP), true negative (TN), false positive (FP) and false negative (FN) concept as described in Table I. Meanwhile, *k*-fold cross validation is employed to verify the performance of the selected model; ResNet-50 in classifying the image as either normal or abnormal (ischemic).

TABLE. I. Assessment Metrics for Classifier Performance

Metrics	Formula	Description
Precision	$\frac{TP}{TP+FP}(\times 100\%)$	Percentage of classified slices that are actually positive
Recall	$\frac{TP}{TP+FN}(\times 100\%)$	Percentage of slices correctly classified
F1-score	$\frac{2TP}{2TP+FP+\ FN}(\times 100\%)$	Percentage of classification performance in term of recall and precision
Specificity	$\frac{TN}{FP+TN}(\times 100\%)$	Percentage of predicting the negative slices
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}(\times 100\%)$	Percentage true or false slices correctly classified

A. Phase 1: Initial Training of Fine-Tuning VGG-16, GoogleNet and ResNet-50

The result of initial training of VGG-16, GoogleNet and ResNet-50 demonstrate average accuracy of 92.2%, 99.4% and 100% respectively as shown in Table II. It can be seen that the VGG-16 in stage-1 training has achieved 86.5% of precision, 92.8% of F1-score and 84.4% of specificity by misclassifying 25 images in the normal class. Meanwhile, GoogleNet in stage-1 training has surpassed the performance of VGG-16 with 98.8% of precision, 99.4% of F1-score and 98.8% of specificity by misclassifying only two images in the normal class. This model closely follows the performance of the ResNet-50 which delivers 100% for all metrics performance. Although misclassified cases have been found in a normal class, good performance has been demonstrated by these three models. All models provide 100% of sensitivity with all false negative cases are correctly classified.

Therefore, to ensure these models are properly trained, they are studied in further detail though validation phase as tabulated in Table III. The average validation accuracy of ResNet-50 is brought up to 100% as compared to VGG-16 and GoogleNet. It also can be observed that VGG-16 in phase-1 validation has yielded 83.3% of precision, 90.9% of F1-score and 80% of specificity by misclassifying eight images in the normal class. GoogleNet has made an improvement in phase-1 validation with 97.6% of precision, 98.8% of F1-score and 97.5% of specificity by misclassifying only one image in the normal class. Among these three models, ResNet-50 have provided the flexibility of being applicable in the case of classifying normal and ischemic PF slices of CT images.

The example of false predicted images in VGG-16 and GoogleNet in stage-1 are shown in Fig. 8. The predicted label shows the PF slices as abnormal while the true class label is normal. The probability for the VGG-16 to detect the PF slice as normal has achieved only 0.0819. In the meantime, the GoogleNet identifies the PF slice as normal with a probability value of 0.4676. These outputs are expected to be similar in colour intensity between the normal and ischemic region of the PF slices.

TABLE. II. TRAINING PERFORMANCE

Models	Precision	Recall	F1-score	Specificity	Accuracy
VGG-16	86.5%	100%	92.8%	84.4%	92.2%
GoogleNet	98.8%	100%	99.4%	98.8%	99.4%
ResNet-50	100%	100%	100%	100%	100%

TABLE. III. VALIDATION PERFORMANCE

Models	Precision	Recall	F1-score	Specificity	Accuracy
VGG-16	83.3%	100%	90.9%	80%	90%
GoogleNet	97.6%	100%	98.8%	97.5%	98.8%
ResNet-50	100%	100%	100%	100%	100%

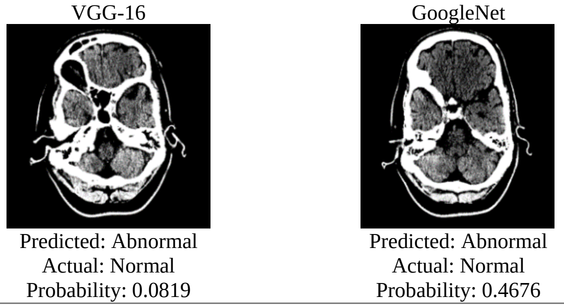


Fig. 8. Incorrect Classified Images after Validation.

B. Phase 2: K-Fold Cross Validation of the ResNet-50 Model

The best model in the phase-1 is retrained using 5-fold cross validation as discussed in Section III (E). Table IV shows the performance results obtained for all the 5-fold cross validation. Similar to the phase-1, the model's performance performs 100% for validation average precision, recall F1-score, specificity and accuracy for normal and ischemic classes. It can be observed that the ResNet-50 model correctly classified and assigned the categories of the images for all folds. The visual representations learned by the ResNet-50 model seems well suited to match the visualization. Fig. 9 demonstrates the feature map of a sample image in various layers of ResNet-50 model. The first layer differs significantly from the other layers as it performs as a group of different edge detectors and almost all the information is preserved by the activations as demonstrated in Fig. 9(b). The model described here may, therefore, keep less information for higher layers and somewhat more abstract as depicted in Fig. 9(c). Further layer in Fig. 9(d) represent the non-existing encoded pattern of the blank filter in the input image.

TABLE. IV. THE AVERAGE PRECISION, RECALL, F1-SCORE, SPECIFICITY AND VALIDATION ACCURACY

Classes	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**Specificity (%)**	**Accuracy (%)**
Normal	100	100	100	100	100
Ischemic	100	100	100	100	

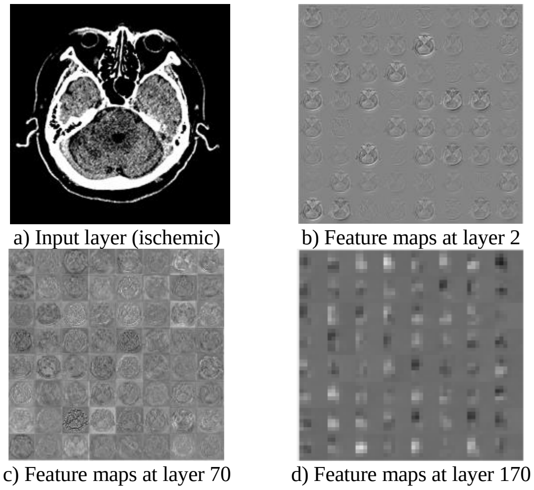

a) Input layer (ischemic)
b) Feature maps at layer 2
c) Feature maps at layer 70
d) Feature maps at layer 170

Fig. 9. Feature Map of Sample Image at different Layers.

C. Processing Speed

The processing speed of the pre-trained models also is evaluated for in training and validation of the PF CT images as tabulated in Table V. Although VGG-16 consists of the least number of layers, poor performance of processing speed has been demonstrated for training and validation. As shown in the results, VGG-16 takes a longer processing speed to train, compared with 6 minutes and 15 minutes training time of GoogleNet and ResNet-50 respectively. In order to improve the accuracy performance of VGG-16, this model requires more epoch which demands a longer processing speed in training. GoogleNet appears to possess relatively good processing speed in comparison to other proposed pre-trained models. The processing speed of ResNet-50 for training and validation demonstrates that the model seems comparable that in GoogleNet model. ResNet-50 only takes around 0.19 seconds on average to classify an image. The ResNet-50 model also requires less epoch to achieve 100% accuracy. This high-processing speed of proposed model may offer some benefits in practical usage in hospitals to assist radiologists in their daily tasks.

D. Comparison to other Methods

There are several studies on DL that have been conducted to address the classification of normal and ischemic brain CT images on the anterior part using different database. Our study is conducted specifically on PF slices. Table VI shows the comparison of the related works on automated classification of the normal and ischemic image using CT image. Pereira et al. [19] use the architecture of CIFAR-10 and ImageNet with the integration of PSO for the classification of normal and ischemic images. They perform their proposed method on 300 CT brain images and have achieved 99% accuracy. Chin et al. [20] develop a CNN model from scratch for the patches classification purpose. They have reported accuracy performance with more than 90% for 256 CT images. Dourado et al. [12] compare the performance of CNN models as the feature extraction with different ML classifiers. They achieved 100% accuracy with 840 CT images. In the above-mentioned state-of-the-art studies, there are two studies which used CNN as feature extraction and classifier. Meanwhile, there is only one study that applies CNN and ML as feature extraction and classifier respectively. In our proposed study, three CNN models are applied and the models performance to classify the normal and ischemic in PF CT slices is observed. The proposed method has achieved 100% accuracy with 5-fold cross validation using 400 images with ResNet-50 model.

The great advantage of the proposed method is that this is the first time the CNN pre-trained-based approach is investigated to use in classifying normal and ischemic mainly in PF slices. The application of pre-trained model as a deep transfer learning offers benefit to address the poor performance of DL using a limited amount of data. Gathering a sufficient amount of data can be a difficult process since there is no existing data for such amount which publicly available to be used for the DL purpose. However, this limitation can be addressed using the data augmentation technique.

TABLE. V. PROCESSING SPEED OF TRAINING AND VALIDATION

CNN models	**Training**	**Validation**	
	1 set	1 set	1 image
VGG-16	33 minutes and 22 seconds	32.9 seconds	0.41 seconds
GoogleNet	6 minutes and 12 seconds	7.1 seconds	0.08 seconds
ResNet-50	15 minutes and 37 seconds	15.4 seconds	0.19 seconds

TABLE. VI. COMPARISON OF WORKS CONDUCTED ON ISCHEMIC STROKE DETECTION USING DL

Author(s)	**Application**	**Number of images**	**Method/architecture uses**	**Accuracy**
Pereira et al. [19]	Anterior slices	300	CIFAR-10 + ImageNet + PSO	Up to 99%
Chin et al. [20]	Anterior slices	256	CNN from scratch	>90%
Dourado et al.[12]	Anterior slices	840	CNN + ML classifiers	100%
The proposed model	PF slices	400	ResNet-50	100%

V. CONCLUSION

In this study, the aim of this experiment is to evaluate three popular pre-trained deep learning CNN architecture; VGG-16, GoogleNet and ResNet-50 to classify normal and ischemic in PF slices of brain CT. In contrast to the usual study of DL in brain CT, the application of DL is studied in further detail for PF slices. An intensive experiment has been conducted on the CT dataset of PF using DL with augmentation technique. Among these three proposed models, ResNet-50 achieves better performance than the rest of the state-of-the-art methods with 100% accuracy for initial training, validation and 5-fold cross validation. The experiment also shows that ResNet-50 is generally applicable to the ischemic classification in PF slices with good processing speed that is less than 1 seconds. Therefore, this model can help to improve the diagnostic performance with better computational efficiency. In future studies, the integration of the detection task with classification can be developed for localizing the location of ischemic in PF slices. A Graphical User Interface (GUI) also will be developed for assisting the diagnosis process in the real-time application.

ACKNOWLEDGMENT

This study is supported by Universiti Tun Hussein Onn Malaysia and UKM Medical Centre under TIER 1 Vot H203 and GPPS Vot U954 Research Grants.

REFERENCES

[1] Aziz ZA, Lee YY, Ngah BA, Sidek NN, Looi I, Hanip MR, Basri HB: Acute stroke registry Malaysia, 2010-2014: results from the National Neurology Registry. Journal of Stroke and Cerebrovascular Diseases. 24(12), 2701-2709 (2015).

[2] Gomez, CR: Time is brain: the stroke theory of relativity. Journal of Stroke and Cerebrovascular Diseases. 27(8), 2214-2227 (2018).

[3] Dubey P, Pandey S, Moonis G: Acute stroke imaging: recent updates. Stroke Research and Treatment. 2013, 6 pages (2013).

[4] Kooi CW, Peng HC, Aziz ZA, Looi I: A review of stroke research in Malaysia from 2000-2014. Med J Malaysia. 71, 58-69 (2016).

[5] Habibi-koolaee M, Shahmoradi L, Niakan Kalhori SR, Ghannadan H, Younesi E: Prevalence of Stroke Risk Factors and Their Distribution Based on Stroke Subtypes in Gorgan: A Retrospective Hospital-Based Study—2015-2016. Neurology Research International. (2018).

[6] El-Koussy M, Schroth G, Brekenfeld C, Arnold M: Imaging of acute ischemic stroke. European Neurology. 72(5-6), 309-16 (2014).

[7] Lin, M P.: Imaging of ischemic stroke. Continuum: Lifelong Learning in Neurology. 22(5):1399 (2016).

[8] Sharon M, Boyle K, Yeung R, Symons SP, Boulos MI, Aviv RI: The predictive value of a targeted posterior fossa multimodal stroke protocol for the diagnosis of acute posterior ischemic stroke. Neurovascular Imaging. 2(1):3 (2016).

[9] Hixson HR, Leiva-Salinas C, Sumer S, Patrie J, Xin W, Wintermark M: Utilizing dual energy CT to improve CT diagnosis of posterior fossa ischemia. Journal of Neuroradiology. 43(5), 346-352 (2016).

[10] Hwang DY, Silva GS, Furie KL, Greer DM: Comparative sensitivity of computed tomography vs. magnetic resonance imaging for detecting acute posterior fossa infarct. The Journal of emergency medicine. 42(5), 559-65 (2016).

[11] Tang FH, Ng DK, Chow DH: An image feature approach for computer-aided detection of ischemic stroke. Computers in biology and medicine. 41(7), 529-36 (2011).

[12] Dourado Jr CM, da Silva SP, da Nóbrega RV, Barros AC, Rebouças Filho PP, de Albuquerque VH: Deep learning IoT system for online stroke detection in skull computed tomography images. Computer Networks. 152, 25-39 (2019).

[13] Tyan YS, Wu MC, Chin CL, Kuo YL, Lee MS, Chang HY: Ischemic stroke detection system with a computer-aided diagnostic ability using an unsupervised feature perception enhancement method. Journal of Biomedical Imaging. 2014, 19 (2014).

[14] Kanchana R, Menaka R: Computer reinforced analysis for ischemic stroke recognition: a review. Indian J. Sci. Technol. 8(35), 81006 (2015).

[15] Kanchana R, Menaka R: A novel approach for characterisation of ischaemic stroke lesion using histogram bin-based segmentation and gray level co-occurrence matrix features. The Imaging Science Journal. 65(2), 124-36 (2017).

[16] Aggarwal N, Agrawal RK: First and second order statistics features for classification of magnetic resonance brain images. Journal of Signal and Information Processing. 3(02), 146 (2012).

[17] Suzuki, K.: Overview of deep learning in medical imaging. Radiological physics and technology. 10(3), 257-73 (2017).

[18] Talo M, Baloglu UB, Yıldırım Ö, Acharya UR: Application of deep transfer learning for automated brain abnormality classification using MR images. Cognitive Systems Research. 54, 176-88 (2019).

[19] Pereira DR, Reboucas Filho PP, de Rosa GH, Papa JP, de Albuquerque VH: Stroke lesion detection using convolutional neural networks. In 2018 International joint conference on neural networks (IJCNN), pp. 1-6. IEEE (2018).

[20] Chin CL, Lin BJ, Wu GR, Weng TC, Yang CS, Su RC, Pan YJ: An automated early ischemic stroke detection system using CNN deep learning algorithm. In 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), pp. 368-372. IEEE (2017).

[21] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J: Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging. 35(5), 1299-312 (2016).

[22] Pak M, Kim S: A review of deep learning in image recognition. In 2017 4th international conference on computer applications and information processing technology (CAIPT), pp. 1-3. IEEE (2017).

[23] He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. IEEE (2016).

[24] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A: Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. IEEE (2015).

[25] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. (2014).

[26] Kamal, H.: Machine learning in acute ischemic stroke neuroimaging. Frontiers in neurology. 9, 945 (2018).

[27] Nouh A, Remke J, Ruland S: Ischemic posterior circulation stroke: a review of anatomy, clinical presentations, diagnosis, and current management. Frontiers in neurology. 5, 30 (2014).

[28] Tsochatzidis L, Costaridou L, Pratikakis I: Deep Learning for Breast Cancer Diagnosis from Mammograms—A Comparative Study. Journal of Imaging. 5(3), 37 (2019).

[29] Khan S, Yairi T: A review on the application of deep learning in system health management. Mechanical Systems and Signal Processing. 107, 241-65 (2018).

[30] Agarwal R, Diaz O, Lladó X, Yap MH, Martí R: Automatic mass detection in mammograms using deep convolutional neural networks. Journal of Medical Imaging. 6(3), 031409 (2019).

[31] Ji Q, Huang J, He W, Sun Y: Optimized Deep Convolutional Neural Networks for Identification of Macular Diseases from Optical Coherence Tomography Images. Algorithms. 12(3), 51 (2019).

[32] Ee CS, Sim KS, Teh V, Ting FF: Estimation of window width setting for CT scan brain images using mean of greyscale level to standard deviation ratio. In 2016 International Conference on Robotics, Automation and Sciences (ICORAS), pp. 1-6. IEEE (2016).

[33] Suberi AA, Zakaria WN, Tomari R, Fuad NF: Classification of Posterior Fossa CT Brain Slices using Artificial Neural Network. Procedia Computer Science. 135, 170-177 (2018).

[34] Lopez AR, Giro-i-Nieto X, Burdick J, Marques O: Skin lesion classification from dermoscopic images using deep learning techniques. In 2017 13th IASTED International Conference on Biomedical Engineering (BioMed), pp. 49-54. IEEE (2017).

Effect of e-Commerce Platforms towards Increasing Merchant's Income in Malaysia

M.Hafiz Yusoff*,[1], Mohammad Ahmed Alomari[2], Nurul Adilah Abdul Latiff[3], Motea S. Alomari[4]

Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA)[1, 2]
Faculty of Ocean Engineering Technology & Informatics, Universiti Malaysia Terengganu (UMT)[3]
Faculty of Engineering, Universiti Putra Malaysia, Selangor, Malaysia[4]

***Abstract*—In 2018, 'Hootsuite' and 'We Are Social' have reported in their Digital Report that Malaysian Internet users had increased to 25.08 million users, representing 79% of Malaysian population. However, some reports indicate that even with the enhancement of digital technology, many merchants are still not using any e-commerce platform and being sceptic about it. With the pervasive increase of the Internet users, many other reports had been published to understand the relationship between e-commerce and the increasing of merchants' income. Therefore, the objective of this research is to study the involvement of Malaysian merchants in e-commerce platforms. A sample of 1060 respondents had been selected randomly across Malaysia to participate in this research by answering a set of survey questions given online. In general, the results show that many merchants have realized the existence of e-commerce and they are familiar with it. However, they have only been utilizing it for purchasing goods and services. On the other hand, a few numbers of Malaysian merchants do engage with some type of e-commerce platforms to operate their businesses. In addition, the research is performed to identify the impact of using e-commerce on Malaysian merchants' income.**

***Keywords*—Income increase; e-commerce; merchant; Malaysia; e-retailing**

I. Introduction

The boom of Internet era accompanied with the Industrial Revolution 4.0 (IR 4.0) rising in the horizon, all that had enabled e-commerce and m-commerce (mobile commerce) to play a vital role in transforming how electronic trading of selling and buying have evolved tremendously [1]. 5G cellular services as well as other similar technologies are expected to have revolutionary effect on e-commerce activities. That will attract users to engage more in e-commerce services and make more buy and selling online. For selling consumers, that will open new perspectives for larger businesses and higher income [2]. It is estimated that e-commerce jobs introduced by companies have been risen by 400,000 within the last ten years, when compared to 140,000 decline in traditional trading; with center jobs pay 31% more than traditional retail jobs in the same area [3].

E-commerce is a process of buying and selling goods and services. It also contains all intermediate process including advertising, promoting, negotiating, ordering, delivering, and also after sales services. E-commerce also includes managing transaction of any digital goods such as Electronic Fund Transfer (EFT) as well as selling and purchasing of electronic news and information of stock prices including classified advertisement [4]. Nowadays, every second means a new product is being introduced and promoted over the Internet when compared to physical premises like stores and shops which has slower trading. The vast technology advancement as well as 4G-5G services has enabled Internet to become a medium for selling and purchasing various types of products, new and old, via e-commerce platforms. Such type of trading offers more speed and ease for the whole process compared to traditional business method. One doesn't have to be present physically or face to face to complete a purchase or deal. E-commerce also has been proven as the best method to operate businesses in this era of globalization where businesses are boundless for everyone [5, 6].

According to a research conducted [7, 8], there are differences on how the men and women uses the Internet. Men uses the Internet less for selling stuff while women showed more entrepreneurial spirit by selling their goods as well as offering services via multiple e-commerce platforms. The reason might be that most women involved as active merchants in e-commerce are housewives trying to make some pocket money for themselves or their families. Most merchants agree that by using e-commerce platforms, they can reach even further in terms of customers' acquisition and distance.

Sales target is always the first thing in doing businesses where a positive cashflow is essential to keep a business running. Thus, by using e-commerce, lesser cost in promoting and advertising can be induced to reach more potential customers when compared to slower sales in traditional method [3]. However, there are various risks in doing transaction over the Internet. First, the risk of encountering a scammer is far more frequent due to the ease of creating fake account. Second risk might be the low security options offered by e-commerce platforms. In order to perform online transaction, one must provide some basic information such as e-mail address and payment account information such as debit or credit card. Although the data privacy policy has been enforced by lawmakers, it is still exposed to hacking threats. These information can relatively easily be accessed by hackers hunting for valuable data over the Internet and then sell them to third party for any unhealthy purpose [9-12].

Malaysia is considered one of the promising countries to get involved deeper in e-commerce activities in future. During recent years in Malaysia, the technology literacy has increased

*Corresponding Authors

rapidly following the big changes in government policy, especially in education where the ministry had spent a lot of budget to adopt digital technology into classrooms. The Ministry of Communication and Multimedia has also enforced low cost for Internet and has encouraged many Malaysian to dive into the digital world. These policies have tremendously affected Malaysian lifestyle especially on how they shop and purchase goods and services [13-15]. Internet has become the largest medium of money transaction globally.

In recent years, Malaysians have been exposed to many entrepreneurship programs which significantly increased their interest to engage in building their own start-up companies. However, this progress has risen a question; how far these emerging business owners are involved with e-commerce. This research is conducted focusing on two main objectives:

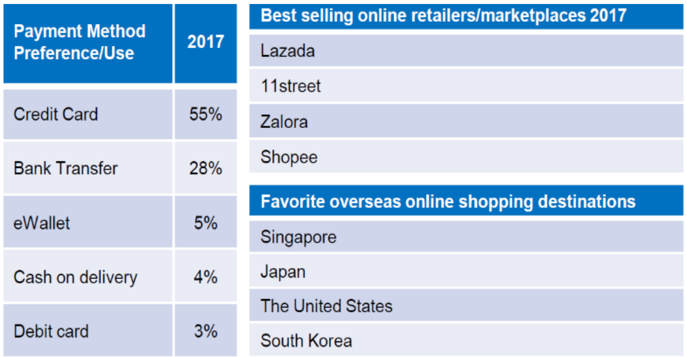
1) To identify the involvement in using e-commerce platforms among Malaysian merchants.

2) To study the impact of using e-commerce platforms towards merchants' income.

The rest of this paper is organized as follows. Details of current research related to this work will be discussed in Section 2. In Section 3, methodology aspects and tools used as well as sampling details will be elaborated. In Sections 4 and 5, the results came out of this study as well as discussion of that results will be presented respectively. Finally, the conclusion of the study will be discussed in Section 6.

II. Literature Review

The rapid growth of e-commerce and m-commerce all over the globe is undeniable. Since the year 2000, the growing adoption of Internet which increased from 52% of total population to 89% in 2018, has boosted e-commerce trading and sales tremendously [16]. According to Ecommerce Foundation 2018 report, out of the 32 million population of Malaysia, the e-commerce shoppers exceeded 70% with high Internet penetration of 82% in this country. In addition, 40% of e-shoppers who purchased online they did that via phone in 2017(Q3). Fig. 1 depicts some details of that e-commerce growth. That is understood with continuous efforts of Malaysian government to facilitate and push e-commerce usage further. In year 2016, the Malaysian Ministry of International Trade and Industry (Miti) has launched a new e-commerce initiative aiming to bring 80% of small and medium-sized companies into the e-commerce world [14].



Payment Method Preference/Use	2017
Credit Card	55%
Bank Transfer	28%
eWallet	5%
Cash on delivery	4%
Debit card	3%

Best selling online retailers/marketplaces 2017
Lazada
11street
Zalora
Shopee

Favorite overseas online shopping destinations
Singapore
Japan
The United States
South Korea

Fig. 1. Malaysia E-Commerce Facts and Figures.

The immense growth of electronic commerce in Malaysia has pushed research community to study that phenomenon as well as problems and challenges facing it. Authors in [17] investigated the effect of online trading on daily life of Malaysians, the development, current status and future trends of online shopping in Malaysia as well as factors that influence e-commerce. They also studied initiatives to raise online shopping by various parties. Ramlan et al. [18] studied the dominant factors that influence Malaysian shoppers to buy online. Usefulness, saving time, and ease of use were the main factors to affect consumers to participate in e-commerce. On the other hand however, privacy, trust, and safety were the main concerns and challenges for e-shoppers. Authors in [4] tried to provide an assessment, evaluation, and understanding of the different aspects on e-commerce in Malaysia. They implemented various tests to compare online consumers versus the offline consumers. Marzuki et al. [15] have used a sample of 160 university students to study how deep they get involved in e-commerce as well as factors behind that. It shows that student involvement is high in e-buying, however it becomes low when it comes to e-business and trading.

E-commerce industry is faced with various challenges. That may include fear of cyber security threats, lack of understanding, limited skills of digital marketing, lack of knowledge regarding market access and regulations in and out of country, and finally lack of experience personnel to run e-commerce activities. Hanefah et al. [19] investigated the tax problems posed by e-commerce and solutions for that problem. Their findings show that tax administration was the main reasons to affect e-commerce growth, followed by double taxation, tax evasion, and tax avoidance. Their results show that subjects perceive e-commerce as creating tax problems.

III. Research Methodology

Various types of research may different type of methods and tools to achieve its target objectives. In this research, a structured questionnaire with descriptive survey has been used to identify the involvement of Malaysians in e-commerce and how e-commerce is affecting merchant's income. The survey was used as a primary data to answer research questions. So, this research is a survey-based research which is intended to identify the impact of e-commerce towards increasing the income of Malaysian merchants. Tuckman [20] stated that survey studies are an efficient way to get information from respondent. It is generally effective way to use descriptive techniques to measure and evaluate the achievement, behavior, perception, and involvement in a specific subject [21].

For this research, a series of 8 questions is distributed via Google Forms. The sample population is a potential Malaysian user of any e-commerce platform. A total of 1300 respondents have completed the questionnaire survey where the responses were screened for errors or incomplete details. Upon the completion of screening process, only 1060 survey forms were considered valid and ready to be entered to data analysis stage. This provides us with a success rate of about 82% which is considered to be appropriate in view of cost and time limitations. The confidence level as well as margin of error used was 95% and 3% respectively. SPSS statistical software version 23 on Windows 10, was used to data and findings in this study. Various statistical techniques were used to analyze data in this study. Frequency Distribution Analysis technique were used to determine the demographic details of the survey respondents. Cross tabulation and the Pearson Chi-Square Test were also used to analyze various relationships in data under study [22]. The group of respondents has been categorized based on sex, age, and occupation. Furthermore, the selected sample consists of respondents from four different jobs within Malaysia, namely students, private workers, government employees, and merchants or self-employed.

IV. Research Findings

In this section, the results obtained from this study will be depicted where the data analysis collected will be presented and discussed.

A. *Respondent Profile*

Table I shows the entire profile of 1060 respondents involved in the survey. They are categorized based on sex, age, and occupation where they are all Malaysians. Table shows that 66% of respondents are women and 34% are men. In this study, most respondents lie within 20-29 years old with about 61% where students are dominating the survey sample by 64% of total respondents.

B. *Awareness of e-Commerce*

Table II depicts the distribution of respondents towards how they are aware of e-commerce. The data conclude that majority of Malaysians are already aware of the existence of e-commerce platforms with more than 82% answered ‘yes’ while only 18% said that they have never heard about e-commerce.

C. *Respondent Status*

Table III identify either the respondent is a merchant or not. Surprisingly, despite being well literate in e-commerce, only 22 respondents (21%) involve as active merchants in any e-commerce platforms, while the balance of 84 respondents (79%) are not online merchant.

TABLE. I. Respondent Profile

Item		**Frequency**	**Percentage**
Sex	Men	362	34.2%
	Women	698	65.8%
Age	19 and below	91	8.6%
	20-29 year old	642	60.6%
	30-39 year old	154	14.5%
	40-49 year old	102	9.6%
	50-59 year old	71	6.7%
	60 and above	0	0%
Occupation	Student	677	63.9%
	Merchant/Self-Employed	102	9.6%
	Government staff	207	19.5%
	Private sector worker	74	7%

TABLE. II. Knowledge of e-Commerce

Do you know anything about e-comerce?		
	Amount	Percentage
Yes	869	81%
No	191	18%

TABLE. III. Respondent Status

Do you practice online business?		
	Amount	Percentage
Yes	224	21.1%
No	836	78.9%

D. *Platforms used to Promote Goods and Services*

Based on Table IV, social media is identified as the most frequently used platforms to promote goods and services. This must be due to many Malaysian spent hours using social media. 63% of respondents use social media as the data shows. Large online markets such as Lazada and Shoppe are also being favored by respondents as much as 19% of total respondents. Falling behind are other platforms; mobile applications 8.6%, self-published websites 8% and others by less than 2%.

E. *Impact of using e-Ccommerce Platforms Towards Increasing Income*

Based on Table V, findings show that 1035 respondents (98%) agree that e-commerce platforms can provide positive impact to increase merchants’ income. Only 2% of respondents disagree of the positive impact. The reason is unknown, and this category might need to be surveyed further to study their negative response.

F. *Impact of using e-Commerce Platforms Towards Increasing Customers*

This research concurs that majority of respondents (95%) agree that e-commerce platforms can enable a merchant to reach more customers; and therefore, increasing the amount of sales and profit. Only 5% of the survey responded by disagreeing as Table VI shows. We were unable to state the reason behind their negative response.

TABLE. IV. Platforms used to Promote Goods and Services

Platform	Frequency	Percentage
Social media	667	62.9%
Self-published website	85	8%
E-commerce market	197	18.6%
Mobile Applications	91	8.6%
Other	20	1.9%

TABLE. V. Impact of using e-Commerce Platforms Towards Increasing Income

Do you agree that e-commerce can increase the merchants’ income?		
	Amount	Percentage
Yes	1035	97.6%
No	25	2.4%

TABLE. VI. IMPACT OF USING E-COMMERCE PLATFORMS TOWARDS INCREASING CUSTOMERS

D you agree that e-commerce platforms towards increasing customers?		
	Amount	Percentage
Yes	1006	94.9%
No	54	5.1%

G. Impact of using e-Commerce Platforms Towards Achieving Higher Sales Target

Table VII shows that more than 90% of respondents agree that e-commerce platforms can significantly help to increase sales target as it enables higher customer reach. However, 9% of respondents disagree of the capability of e-commerce platforms to increase sales target.

TABLE. VII. IMPACT OF USING E-COMMERCE PLATFORMS TOWARDS ACHIEVING HIGHER SALES TARGET

Do you agree that e-commerce is able to increase sales target?		
	Amount	Percentage
Yes	962	90.8%
No	98	9.2%

H. Suggestion to Promote using e-Commerce

Finally, the research intended to know if the respondents would suggest any e-commerce platforms to the people around them. Results in Table VIII shows that 83% of them will suggest it to their friends and family while 17% are not interested to do that.

TABLE. VIII. SUGGESTION TO PROMOTE USING E-COMMERCE

Would you suggest any e-commerce platforms to your friends and family?		
	Amount	Percentage
Yes	884	83.4%
No	176	16.6%

V. DISCUSSION

In recent years, e-commerce and m-commerce have influenced many Malaysians to participate either as consumers or retailers. Based on this research finding, it shows that Malaysian women are more responsive to the survey conducted than men. The big range of age (19 – 60 years old) proves that senior citizens are also actively engaged with digital transformation of businesses. This is a good sign that Malaysians regardless of age and generations have accepted and adopted the digital technology in their daily life. Findings also show that Malaysians are aware of the existence of e-commerce platforms. Many reports state that Malaysians are one of most active buyers in South East Asia region, but not as a merchant or trader. This research verifies that statement by identifying that 20% out of the 1060 respondents are not doing any form of businesses online. This research also shows that most respondents use social media (62%) as their main platform to promote products online. This indicates that many Malaysians still don’t know how to use proper e-commerce platforms to create their own online shops.

By using e-commerce platforms, respondents agree that e-commerce is significantly helping to increase merchants’ income. However, the findings found some contradiction since there is small incline on positive impact to increasing customers and sales target. This might be due to bad experience using e-commerce platforms or using unsuitable platforms for their goods and services. Practicing business require a set of skill while using e-commerce platforms to gain more reachability and increase. It is not easy to transform from traditional business method in order to adopt e-commerce platforms. The strategy is totally different between the two methods. Due to the depth of understanding, some merchants did profit from e-commerce platform while some did not. Most probably that might be due to the nature of some businesses which require face to face interaction or to be physically present to make it happen. There are many pros and cons in using e-commerce. It is proven that e-commerce is prone to have some security issues. Additionally, the tendency to encounter scammers are higher in e-commerce than practicing traditional businesses. Nonetheless, data shows that most Malaysians are willing to suggest using e-commerce to their friends and family.

VI. CONCLUSION

Although E-commerce is considered new to Malaysia, it is picking up fast as the Malaysian government enforces new policies and adapting to Industrial Revolution 4.0. The rising of lots of start-ups also contributes to the emerging of e-commerce platforms used in Malaysia which make more Malaysians to become entrepreneurs every day. However, the usage of proper e-commerce platforms among Malaysian merchants are way too low compared to the number of consumers. In order to encourage the digital transformation of merchants, more initiatives need be done such as developing safer and more secure e-commerce platforms. More education programs on e-commerce should also be available in order to attract more merchants to transform and digitize their businesses. A couple of conclusions can be made based on this research. First, e-commerce is able to increase customer acquisition, sales target and income. Second, Malaysians are active users of e-commerce despite the various threats encountered during the digital business. Furthermore, many more research need be conducted to identify factors that affect e-commerce suitability for Malaysian merchants during new services supported by IR 4.0 and 5G cellular technology.

REFERENCES

[1] Chao, Chiang-nan. "Emergence Impacts of Mobile Commerce: An Exploratory Study." Journal of Management and Strategy 8.2 (2017): 63-70.

[2] N. Kshetri, "5G in E-commerce activities," IT Prof, vol. 20, no. 4, pp. 73-77, 2019.

[3] M. Mandel, "How Ecommerce Creates Jobs and Reduces Income Inequality," Progressive Policy Institute, 2017.

[4] A. C. P. Harn and A. K. d. H. b. Ismail, "ECommerce: A Study on Online Shopping in Malaysia," Journal of Social Sciences 13.3 pp. 231-242, 2006.

[5] R. Tiwari, S. Buse, and C. Herstatt, "From electronic to mobile commerce: technology convergence enables innovative business services," Hamburg University of Technology (TUHH), 2006.

[6] Garín-Muñoz, Teresa, et al. "Models for individual adoption of eCommerce, eBanking and eGovernment in Spain." Telecommunications Policy 43.1 (2019): 100-111.

[7] S. Rodgers and M. A. Harris, "Gender and e-commerce: An exploratory study," Journal of advertising research, vol. 43, no. 3, pp. 322-329, 2003.

[8] C. V. Slyke, F. Belanger, and R. Hightower, "Understanding gender-based differences in consumer e-commerce adoption," in Proceedings of the 2005 Southern Association of Information Systems Conference, 2005.

[9] A. S. H. M. Rawi, S. Z. Omar, and M. S. S. Ali, "Level of relationship between various selected factors with the intention to use e-commerce among Internet users," Malaysian Journal of Media Studies, vol. 13, no. 2, pp. 11–28, 2011.

[10] Lee, M. Factors influencing the adoption of internet banking: An integration of TAM & TPB with perceived risk & perceived benefit. Electronic Commerce Research & Applications, 8(3), 130–141, 2009.

[11] Ghosh, Anup K., and Tara M. Swaminatha. "Software security and privacy risks in mobile e-commerce." Communications of the ACM 44.2 (2001): 51-57.

[12] Singh, Sachchidanand, and Nirmala Singh. "Internet of Things (IoT): Security challenges, business opportunities & reference architecture for E-commerce." 2015 International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE, 2015.

[13] Hootsuite, "Digital in 2018 Report", Available from: https://hootsuite.com/pages/digital-in-2018, [Accessed 15 February 2019].

[14] Ecommerce Foundation, "2018 Global Ecommerce Report", https://www.internetalliance.my/wp-content/uploads/ 2018/10/Global-B2C-e-Commerce-Country-Report-2018.pdf, [Accessed 26 May 2019].

[15] N. Mat, N. Marzuki, J. Alias, and N. A. Abdullah, "Student Involvement in E-Commerce: A Case Study in UKM," Malaysian Journal of Student Advancement, vol. 19, no. 2, pp. 59-69, 2016.

[16] NationMaster, Online Shopping Trends 2019 & Key Figures - What you need to know, https://www.nationmaster.com/ ecommerce.

[17] Paynter, John, and Jackie Lim. "Drivers and impediments to e-commerce in Malaysia." Malaysian Journal of Library & Information Science 6.2 (2001): 1-19.

[18] R. Ramlan and F. Z. Omar, "A study on factor that influence online shopping in Malaysia," in 5th International Conference of the Asian Academy of Applied Business (AAAB), 9th - 10th June 2011, Cambodia, 2011.

[19] Hanefah, Hajah Mustafa Mohd, Haslinda Hassan, and Zaleha Othman. "E-commerce Implications Potential Problems and Challenges in Malaysia." International Business Research 1.1 (2008): 43-57.

[20] B. W. Tuckman, Conducting Educational Research. Fort Worth: Harcourt Brace College Publishers, 1990.

[21] W. Wiersma, Research Methods in Education: An Introduction. Boston Allyn and Bacon, 1995.

[22] White, D., and A. Korotayev. "Statistical analysis of cross-tabs." Anthrosciences. org (2004).

Novel Adaptive Auto-Correction Technique for Enhanced Fingerprint Recognition

Thejaswini P[1]
Department of ECE
JSS Academy of technical education
Uttarhalli-Kengeri Road, Bangalore
560060, India

Srikantaswamy R S[2]
Department of ECE
Siddaganga Institute of Technology
BH Road, Tumkur 572103
India

Manjunatha A S[3]
Department of CSE
Siddaganga Institute of Technology
BH Road, Tumkur 572103
India

***Abstract*—Fingerprints are the most used biometric trait in applications where high level of security is required. Fingerprint image may vary due to various environmental conditions like temperature, humidity, weather etc. Hence, it is necessary to design a fingerprint recognition system that is robust against temperature variations. Existing techniques such as automated and non-automated techniques are not real time analysis (adaptive). In this paper, we propose an adaptive auto correction technique called Reference Auto-correction Algorithm. This proposed algorithm corrects user reference fingerprint template automatically based on captured fingerprint template and the matching score obtained on daily basis to improve the recognition rate. Analysis is carried out on 250 fingerprint templates stored in the database of 10-users captured at varying temperature from 25^{0}C to 0^{0}C. The experimental result shows 40% improvement in the recognition rate after applying auto correction algorithm.**

***Keywords*—Minutiae, Euclidean distance; artificial neural network; CN- crossing number; reference auto-correction; adaptive method; ISO template; auto-correction algorithm**

I. Introduction

Advancement in the technology has led biometrics to replace the conventional access control and time attendance systems [1]. Biometric systems [2] are used for recognition of user using various biometric traits like Fingerprints, IRIS, Face, Voice, DNA etc. These biometric traits are very unique features of human beings and usually remain stable throughout the life span under normal conditions. Error rate implications, accuracy and response time of fingerprint are better when compared to other biometric traits. Thus the fingerprint develops the obligatory variable in striking security and reliable empathy of the individual. Fingerprints are utilized as means of security in places such as voting, banking operations, day to day attendance system etc. The quality of fingerprint image varies on many characteristics such as humidity, weather, temperature etc., which affects the performance of the biometrics system. Even though fingerprints are considered as very stable and unique throughout the life, it has been observed that the fingerprint pattern varies due to environmental variations like temperature, humidity, dust and also due to ageing of the person [3].

Fingerprint based biometric system works on the principle of comparing live fingerprint image with the reference fingerprint image stored in the database to find matching [4]. The variation of fingerprint image will affect the quality of captured input fingerprint image leading to poor matching with the stored reference fingerprint image in the database and hence it fails to recognize the same person’s identity [5] [6]. This causes lots of inconvenience to use the fingerprint based biometric system for recognition of users especially in day to day time attendance system [7]. In spite of optimal noise reduction using filters, the fingerprint image captured may fail to match with the reference fingerprint image stored in the database due to environmental variation (like temperature), which varies the features in the captured fingerprint image. The issue of rejection of fingerprint image of some user due to variation of environmental parameter like temperature must be addressed.

An attempt is made to improve the recognition rate of fingerprint image even though the matching of fingerprint image fails after applying optimal filters under the environmental variations (like temperature variations). Hence to deal with the failure to recognize the fingerprint image of the user under the environmental variation (like temperature variation), a study and analysis of the fingerprint images are carried out. A novel auto-correction algorithm for autocorrecting the reference fingerprint template (template is the extracted minutiae features of the fingerprint image) without manually re-registering the fingerprint image of the user is proposed so that, the fingerprint recognition system is robust against the changes due to environmental parameters like temperature. Organization of sections in the paper is as follows: Section II demonstrates the proposed reference auto-correction method for fingerprint images, Section III discuss the results and experimental analysis of different datasets for users and Section IV concludes the research outcome.

II. Proposed Fingerprint Recognition and Reference Auto-Correction Technique

Due to environmental changes in the real world, fingerprint pattern may be altered. Even little variations in the fingerprint pattern leads to varied matching score in case of the authentication system, thus resulting in rejection of genuine user. The proposed reference auto-correction algorithm mainly focuses on fingerprint based time-attendance system, where the user has to place the finger daily on the fingerprint sensor to provide fingerprint image for verification. The principle behind auto-correction algorithm is, to correct user reference fingerprint template automatically based on the everyday captured fingerprint template and the matching score obtained. Hence, recognition rate of authentication system can be improved.

The proposed Fingerprint Recognition and Reference Auto-Correction technique comprises of four stages:

- Minutiae Feature Extraction
- Euclidean distance Calculation
- Adaptive ANN
- Reference Auto-Correction

First, extract the minutiae features from the captured fingerprint image and store in the database as Reference Fingerprint Template (RFPT). For verification or authentication of user, the extracted fingerprint template from captured fingerprint image is compared with the reference fingerprint template stored in database. The Euclidean distance calculation is used for obtaining matching score based on the similarities [8] between two fingerprint templates. If the matching score is greater than the value of threshold (Th), the user is recognized (success or accepted). In such case, the captured fingerprint templates are stored in the cache database. The cache database is implemented as FIFO with the size of 30 templates (i.e. to store previous 30 days fingerprint templates). After 30 templates, the first stored template is deleted from cache. At any given point of time, previous 30 days templates are available in the cache for any user.

In the process of verification, if the matching score is less than the value of Threshold (Th), then the user is not recognized. In such cases, the user fingerprint template might have been changed beyond the threshold limit and hence failed to recognize. This situation may occur over a period of time due to environmental parameter like temperature variation, which leads to the variation of features in the captured fingerprint image due to introduction of noise.

The proposed flowchart of Fingerprint Recognition and Reference Auto-Correction system is shown in Fig. 1.

A. Minutiae Feature Extraction

Minutiae are the most widely used feature for representing the fingerprint due to its virtues compared with other features [9]. The minutiae features are extracted through alignment and pairing. Here, the essential minutiae features are extracted through the Crossing number (CN) identification approach. In CN minutiae feature extraction method, the orientation of minutiae point together with its location is obtained by using eight connected pixels for all the fingerprints.

For a pixel 'b', the crossing number can be obtained as:

$$Num_{Cr} = \frac{1}{2}\sum_{n=1}^{8}\left|b(n) - b(n+1)\right| \tag{1}$$

Where, $b(9) = b(1)$

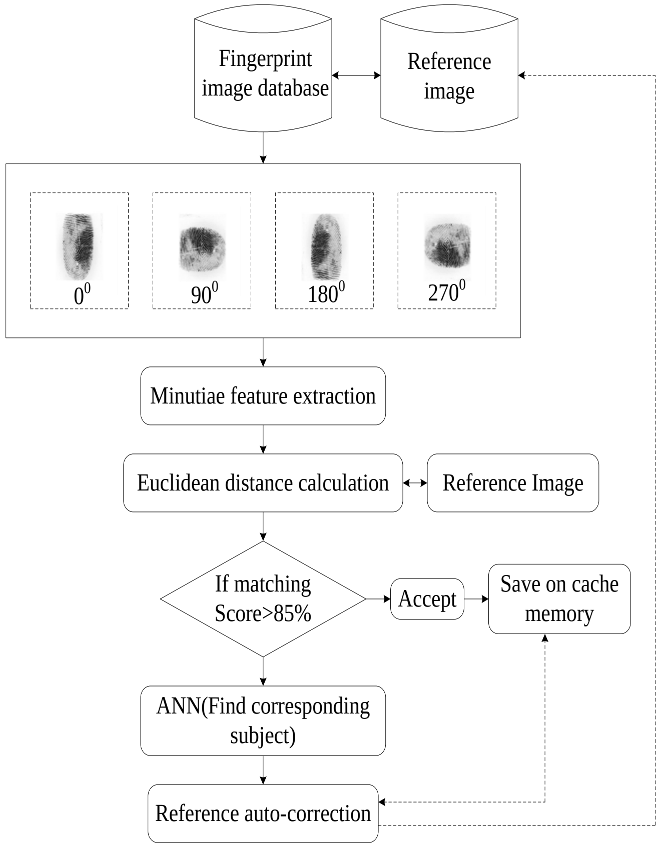

Fingerprint image database
Reference image
0^0
90^0
180^0
270^0
Minutiae feature extraction
Euclidean distance calculation
Reference Image
If matching Score>85%
Accept
Save on cache memory
ANN(Find corresponding subject)
Reference auto-correction

Fig. 1. Fingerprint Recognition and Reference Auto-Correction System.

In the above equation, $b(n)$ is the neighbor pixel of 'b' with ($b(n) = 0/1$). By scanning the entire fingerprint image, we can get the minutiae points based on the characteristic value of CN. On applying this algorithm, border area may be ignored. Since the border area of the image will give more false minutiae points [10], there is no need to extract the minutiae points. Once the essential fingerprint minutiae points are extracted, the similarity between any two fingerprints could be found using Euclidean distance calculation.

B. Euclidean Distance Calculation

Euclidean Distance (ED) is a metric used for calculating the similarity between two fingerprint images [11]. In this step, matching score between the RFPT and user input fingerprint image is calculated based on Euclidean Distance calculation. The ED is a well-known distance measure, also referred as Pythagorean distance, which is normally defined as the square root of the sum of the squares of the distance between the corresponding coordinates of two images; or simply as the straight-line distance between two images in the Euclidean space.

The Euclidean distance between the fingerprint images can be computed by means of the following representation:

$$ED_{pq} = \sqrt{\left(m_p - m_q\right)^2 + \left(n_p - n_q\right)^2} \tag{2}$$

Where,

(m_p, n_p), (m_q, n_q) are the pixel coordinates of two fingerprint images

In the process of verification, if the matching score is less than the value of Threshold (Th), it is considered as user is not recognized. In such cases, the user fingerprint template might have been changed beyond the threshold limit and hence failed to recognize. This situation may occur over a period of time due to environmental parameter like temperature variation, which leads to the variation of features in the captured fingerprint image due to introduction of noise.

C. Adaptive ANN

The genuineness of the rejected input fingerprint image during Euclidean distance calculation is checked for the cause of rejection through the Adaptive ANN technique [12] before applying correction on the stored fingerprint templates in the cache database. This is because; rejection can happen when some intruder places the finger instead of genuine user on the fingerprint sensor. In such case, when the rejection happens due to intruder, the correction should not be applied on the stored reference fingerprint template

D. Artificial Neural Network

ANN's are information-processing and computing method stimulated by biological neuron processing [13]. Fingerprint templates in the cache database of the user and the extracted template of input rejected fingerprint images are trained with the ANN by their specific minutiae features. The training is performed in the neural network system on known input and output data by methods of weight alteration until the system can reasonably address the input and output space. ANN contains three layers such as an input layer, hidden layer and output layer. The input layer contains the current information of the network (i.e., the minutiae features of the fingerprints) and the output layer gives the response of the given input (i.e. the class of the fingerprint). In case the obtained output does not facilitate the goal (0=Genuine and 1=Intruder) inside a predefined acceptable error, the weights and inclinations inside the system are adjusted by constraining an error function (generally error function is least of MSE) relating the simulated and the target output. The training technique will be done when the mismatch between simulated and target output is nearly small. ANN classifies the rejected input fingerprint image as genuine user or not at the time of matching.

E. Back- Propagation Algorithm

The backpropagation algorithm starts with the comparison of output pattern with the target vector [14]. The error values are calculated from the hidden units. The changes of incoming weight can begin with the output layer and pass through the hidden layer. The advantage of this algorithm is, it corrects the network weights and decreases the training error of the fingerprint recognition system.

Step 1: Set the initial value of every interconnection weight between the input to hidden and hidden to output layers as a small random number.

Step 2: Import the learning sample pair (i.e. the input minutiae points of the fingerprint samples and its corresponding target with the genuine or fake label) and operate steps 3 to 5 with each sample pair.

Step 3: For every fingerprint minutiae point sets, calculate the output (i.e. genuine or fake fingerprint) of every network layer based on below equation.

$$Y_{(m)}^{NN(out)} = \beta_m + \sum_{x=1}^{X} W_{xm} Y_{(x)}^{NN} \tag{3}$$

Where

$$Y_{(x)}^{NN} = \frac{1}{1 + \exp^{-\left(\sum_{q=1}^{Q} W_{qx} Y_{(q)} + \omega_q\right)}} \tag{4}$$

In the above equation (3), $Y_{(m)}^{NN(out)}$ is the output of the network from m^{th} output node; β_m represents the bias of output node 'm';W_{xm} is the interconnection weight between hidden and output nodes, and $Y_{(x)}^{NN}$ is the response from hidden layer.

Also, in eq. (4), W_{qx} is the interconnection weight between input and hidden nodes; ω_q represents the bias of hidden node 'q';$Y_{(q)}$ represents the input minutiae points.

Step 4: Calculate training error ($\varepsilon_{(m)}^{BP}$) using below equation

$$\varepsilon_{(m)}^{BP} = Y_{(m)}^{NN(target)} - Y_{(m)}^{NN(out)} \tag{5}$$

Where $Y_{(m)}^{NN(target)}$ is the target label of the fingerprints.

Step 5: Correct weights for the next iteration ($W(t+1)$) based on the back propagation error ($\varepsilon_{(1)}^{BP}$) and the weights of the current iteration ($W(t)$).

$$W(t+1) = W(t) + \lambda Y_{(1)} \varepsilon_{(1)}^{BP} \tag{6}$$

In the above equation, λ is the learning rate, the value of the learning rate will be between 0.2 to 0.5.

Step 6: Check whether the outputs meet the accuracy requirement in the process of training every fingerprint templates stored in the cache and terminates the training process.

In the proposed Adaptive ANN, the weights are optimally selected by means of the Fruit fly optimization algorithm (FOA).

F. Fruit Fly Optimization Algorithm (FOA)

Fruit fly optimization algorithm [15] is stimulated by the characteristic fruit flies found in the environment. It is one of the optimization techniques for finding the solution for the input random population. Here, the input population is the random weights of the ANN. The smell foraging phase enables an individual to search and locate food sources around the fruit fly swarm [15]. The fruit fly is better than different species in vision and osphresis. Smell concentration is the fitness value for each food source. Here, the fitness value is the minimum error of the ANN. For each of the food sources, the smell fixation that relates to the fitness esteem is assessed next. Weight values with maximum smell concentration value are allocated in the vision foraging phase.

Accordingly, the Fruit fly goes across that food direction. The search behavior of Fruitfly for optimal weight selection is shown in Fig. 2 and the steps are explained as follows:

1. Initiate the location of fruit fly swarm randomly from X_axis and Y_axis. This is an initial population to update parameter (i.e. weights of NN).

2. To form osphresis search, random number of food sources (i.e. weights) are created around the fruit fly swarm represented by the equations,

$$X_j = X\_axis + Randomvalue(M_p) \tag{7}$$

$$Y_j = Y\_axis + Randomvalue(M_p) \tag{8}$$

Where, (M_p) is the fly range.

3. Calculate the distance and smell concentration judgment value (i.e. fitness value) for finding optimal weights. Here, the fitness value is the minimum error of the Neural Network.

4. Among the fruitfly swarm, find out the fruit fly with maximum smell concentration (i.e. optimal weights when an error in the NN output is minimum).

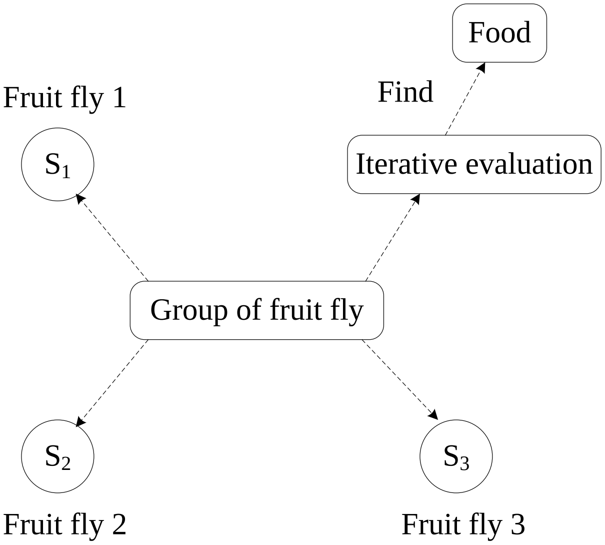


Fig. 2. Search behavior of Fruitfly for Optimal Weight Selection.

5. The swarm direct towards the location which has the best smell (optimal weights) using vision-based search.

6. Stop the algorithm when the iteration number reaches the maximum iteration number otherwise repeat the step 2

The Pseudo code of FOA is as follows (Fig. 3):

```
Input: Randomly initialized interconnection weights
Output: Best smell concentration (i.e. optimal interconnection weights)
// Initialization
Initiate iteration number represented as $I_x$ and population size $P_x$ equivalent to size of weight parameter
// Initiate the swarm location (SL)and fly range ($M_p$)
Iter = 0;
$X\text{-}axis = R(SL), Y\text{-}axis = R(SL);$ where, $R$=random value;
Repeat
While i=1,2,…..$I_x$;
//Osphresis foraging phase
//Generate new weights from random flight direction and distance around current fruit fly swarm location
$X_j=X\_axis+R(M_p);$
$Y_j = Y\_axis + R(M_p);$ where, $M_p$= fly range;
//Calculate Distance
$Dist_j = \sqrt{(X_j^2 + Y_j^2)}$;
// calculate the judgment value for smell concentration
$S_j = 1/Dist_j$;
// calculate smell concentration
$Smell_j = function(S_j)$;
//Identify maximum smell concentration of fruit fly among the whole swarm
$[bestsmell \quad bestindex] = \max(Smell_j)$;
//Vision based search
If ( $smellbest < bestsmell$ )
Return $smellbest = bestsmell$;
$X\_axis = Y\_axis = bestindex$;
$Iter = Iter + 1;$
Until $Iter = I_x$
```

Once optimal weights are found, the same weights are used during the training and testing phases of ANN

G. Reference Auto-Correction

Adaptive Reference Auto-correction method is proposed to automatically correct the reference fingerprint template when the rejection of genuine fingerprint image occurs. Auto-correction of reference fingerprint images are carried out when the genuine fingerprint gets rejected due to environmental variations. The proposed reference auto-correction algorithm is as shown in Fig. 4.

In this technique first, consider the cache of user/subject where the corresponding fingerprint templates are stored on daily basis. By applying Euclidean distance measurement technique on the reference fingerprint template and the stored fingerprint templates in the cache obtain matched and unmatched minutiae points of the corresponding fingerprint templates. Now, separate the matched and unmatched minutiae points of all the fingerprint templates stored in cache database and average all the separated unmatched points to obtain the new minutiae points (unmatched minutiae points are the new features obtained due to variation of temperature). Add these processed minutiae points into the last fingerprint template stored in cache. Resultant fingerprint template is a new template which is the combination of minimum good points of the reference fingerprint template required for matching plus the processed varied minutiae points. Now replace the reference fingerprint template by this new template. So, for next fingerprint image matching, this new template acts as reference template for the user; hence, reducing the percentage of false rejection of genuine subject/user. If the user is not found to be genuine, the rejected input fingerprint image is ignored and not stored in the cache database, nor used for correction, declared as intruder and ignored.

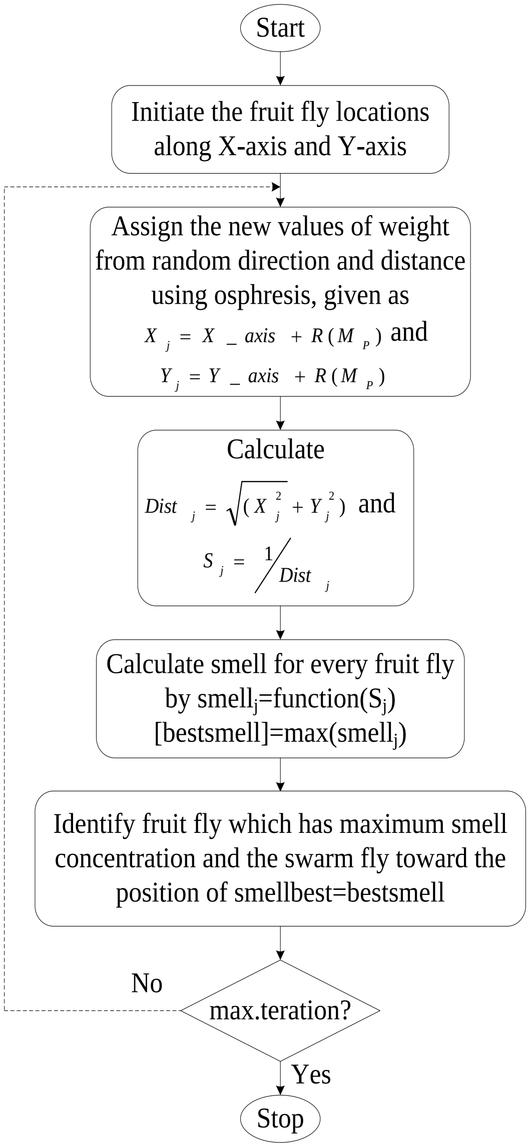

Start
Initiate the fruit fly locations along X-axis and Y-axis
Assign the new values of weight from random direction and distance using osphresis, given as
$X_j = X\_axis + R(M_P)$ and
$Y_j = Y\_axis + R(M_P)$
Calculate
$Dist_j = \sqrt{(X_j^2 + Y_j^2)}$ and
$S_j = 1/Dist_j$
Calculate smell for every fruit fly by smell$_j$=function(S$_j$)
[bestsmell]=max(smell$_j$)
Identify fruit fly which has maximum smell concentration and the swarm fly toward the position of smellbest=bestsmell
No
max.teration?
Yes
Stop

Fig. 3. Flow Diagram of Fruit fly Algorithm.

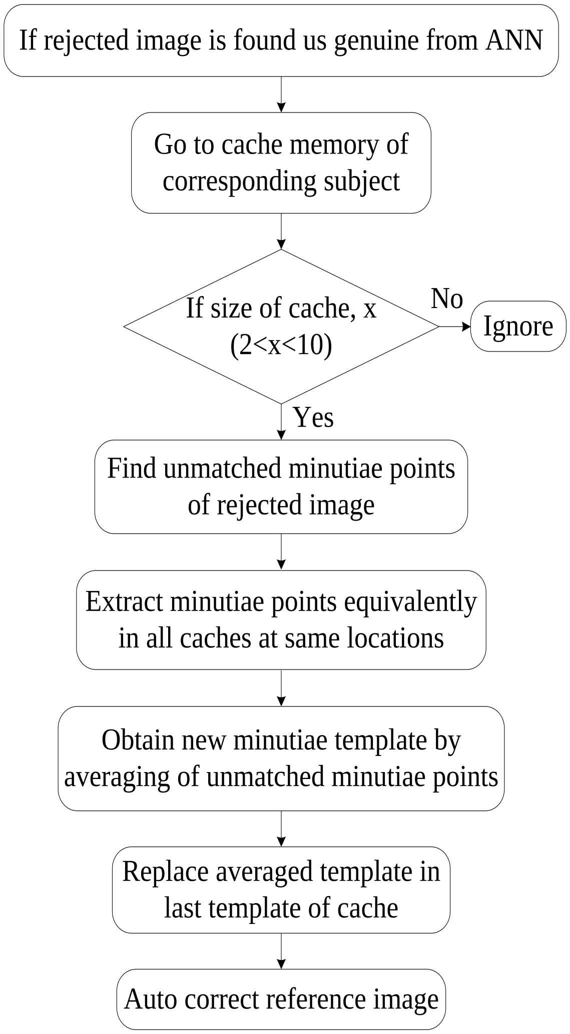

If rejected image is found us genuine from ANN
Go to cache memory of corresponding subject
If size of cache, x (2<x<10)
No
Ignore
Yes
Find unmatched minutiae points of rejected image
Extract minutiae points equivalently in all caches at same locations
Obtain new minutiae template by averaging of unmatched minutiae points
Replace averaged template in last template of cache
Auto correct reference image

Fig. 4. Reference Auto-Correction Algorithm.

III. Result and Discussion

Adaptive reference auto-correction algorithm is used to correct the reference fingerprint template to match with the captured input fingerprint image when the degradation of fingerprint image occurs due to environmental variations (mainly temperature variation). The algorithm is implemented on the real-time fingerprint templates to test the improvement in the recognition of fingerprint templates. The results obtained after applying auto-correction algorithm are compared with the results obtained from without auto-correction. The values are tabulated and graphs are plotted to evaluate the performance.

The experiment is conducted with 10-user/subject fingerprint data which are captured at varying temperature from 25^0C to 0^0C.

First, all 10 users/subjects fingerprint images are captured at room temperature to extract minutiae features and stored as reference fingerprint template for the respective users/subjects in the database. Next, fingerprint images are captured for every user/subject by varying temperature from 25^0C to 0^0C using a closed chamber of controlled temperature. There are about 26-fingerprint images captured for every user/subject with the variation of 1^0C from one fingerprint image to other fingerprint image. The live finger of the user/subject is exposed to a minimum of 3-minutes every time when the temperature is set to a new value.

A. *Analysis*

Analysis is performed on the fingerprint data set stored in the database. Following parameters used in the process,

1) Type of sensor used to capture fingerprint image: Optical, with 500 DPI, 256 gray scales

2) Fingerprint template size- 1x996 bytes

3) Fingerprint image size- 256x256 bytes

4) Template type: ISO-1974-2 format

5) Threshold value set- 85% of 996 minutiae points (i.e., 847 out of 996 minutiae points are required minimum for matching between reference fingerprint template and captured input fingerprint image to be considered as recognized).

The analysis is carried out by comparing the captured fingerprint templates with reference fingerprint template for the respective user/subject using Euclidean distance matching algorithm. The matching is performed without and with using auto-correction algorithm for all the 10-users/subjects fingerprint templates stored in database for varying temperatures from 25^0C to 0^0C and the matching scores are recorded and tabulated as shown in the Table I and Table II. The graphs are plotted showing the variation of matching score against the temperature for both the cases, with and without auto-correction based on the values recorded in the tabular columns are as shown in Fig. 5 and Fig. 6. Analysis of performance of all the users is tabulated in Table III. From Table III it can be observed that, without auto-correction only 2-users fingerprint images are matched the references images of their own. After applying auto-correction the result shows that, 6-users fingerprint images are matched the references images of their own. Hence there is an improvement in the success rate.

In practical scenario, the temperature variations are based on the seasons and also on the regions. The temperature is not varying always from 25^0C to 0^0C. Based on the regions, temperature variations may be few degrees like 20^0C, 18^0C, 15^0C, 12^0C, 10^0C, 8^0C, 6^0C etc. Considering the regional variations in temperature, the results are analyzed for various temperature ranges and tabulated in Table IV. Fig. 7 shows the graph plotted for the values in Table IV.

TABLE. I. MATCHING SCORES OBTAINED BEFORE APPLYING AUTO-CORRECTION ALGORITHM

Temperature in ^{0}c	**Matching Scores**									
	User1	***User2***	***User3***	***User4***	***User5***	***User6***	***User7***	***User8***	***User9***	***User10***
25^0	1	1	1	1	1	1	1	1	1	1
24^0	0.965863	0.950803	0.944779	0.925703	0.954819	0.962851	0.899598	0.952811	0.943775	0.926707
23^0	0.963855	0.937751	0.949799	0.930723	0.954819	0.963855	0.900602	0.970884	0.938755	0.924699
22^0	0.943775	0.939759	0.940763	0.921687	0.940763	0.956827	0.88755	0.9749	0.934739	0.919679
21^0	0.964859	0.939759	0.943775	0.925703	0.938755	0.961847	0.888554	0.965863	0.941767	0.921687
20^0	0.950803	0.934739	0.944779	0.916667	0.933735	0.962851	0.863454	0.962851	0.938755	0.928715
19^0	0.935743	0.921687	0.944779	0.923695	0.935743	0.958835	0.863454	0.959839	0.934739	0.908635
18^0	0.938755	0.915663	0.940763	0.903614	0.929719	0.956827	0.829317	0.965863	0.929719	0.906627
17^0	0.942771	0.913655	0.935743	0.893574	0.927711	0.953815	0.839357	0.943775	0.928715	0.907631
16^0	0.942771	0.893574	0.932731	0.873493	0.921687	0.954819	0.818273	0.954819	0.939759	0.90261
15^0	0.923695	0.878514	0.934739	0.878514	0.923695	0.951807	0.821285	0.948795	0.938755	0.911647
14^0	0.927711	0.866465	0.929719	0.868473	0.925703	0.949799	0.829317	0.953815	0.925703	0.891566
13^0	0.920683	0.858433	0.927711	0.858433	0.914659	0.945783	0.812249	0.941767	0.923695	0.891566
12^0	0.919679	0.851405	0.917671	0.863453	0.907631	0.940763	0.811245	0.934739	0.914659	0.890562
11^0	0.904618	0.843373	0.914659	0.853413	0.909639	0.947791	0.793173	0.934739	0.925703	0.859438
10^0	0.918675	0.841365	0.913655	0.843373	0.899598	0.947791	0.788153	0.932731	0.911647	0.865462
9^0	0.924699	0.823293	0.915663	0.838353	0.901606	0.940763	0.774096	0.932731	0.88755	0.854418
8^0	0.905622	0.80522	0.908635	0.848393	0.90261	0.932731	0.778112	0.933735	0.898594	0.850402
7^0	0.916667	0.799196	0.90261	0.833333	0.896586	0.935743	0.781124	0.925703	0.883534	0.839357
6^0	0.90261	0.783136	0.88755	0.851405	0.891566	0.933735	0.774096	0.927711	0.896586	0.840361
5^0	0.87249	0.774096	0.883534	0.798192	0.891566	0.923695	0.774096	0.911647	0.873494	0.824297
4^0	0.873494	0.777108	0.874498	0.781124	0.891566	0.919679	0.78012	0.914659	0.85241	0.830321
3^0	0.844378	0.763052	0.812249	0.813253	0.86747	0.918675	0.764056	0.884538	0.861446	0.825301
2^0	0.832329	0.758032	0.806225	0.763052	0.862449	0.911647	0.763052	0.906627	0.838353	0.811245
1^0	0.84739	0.744979	0.768072	0.742971	0.772088	0.907631	0.758032	0.87751	0.820281	0.799197
0^0	0.803213	0.740963	0.75	0.758032	0.768072	0.883534	0.74498	0.866466	0.754016	0.801205

TABLE. II. MATCHING SCORES OBTAINED AFTER APPLYING AUTO-CORRECTION ALGORITHM

Temperature in ^{0}c	**Matching Scores**									
	User1	**User2**	**User3**	**User4**	**User5**	**User6**	**User7**	**User8**	**User9**	**User10**
25^0	1	1	1	1	1	1	1	1	1	1
24^0	0.965863	0.950803	0.944779	0.925703	0.954819	0.962851	0.899598	0.952811	0.943775	0.926707
23^0	0.963855	0.937751	0.949799	0.930723	0.954819	0.963855	0.900602	0.970884	0.938755	0.924699
22^0	0.943775	0.939759	0.940763	0.921687	0.940763	0.956827	0.88755	0.9749	0.934739	0.919679
21^0	0.964859	0.939759	0.943775	0.925703	0.938755	0.961847	0.888554	0.965863	0.941767	0.921687
20^0	0.950803	0.934739	0.944779	0.916667	0.933735	0.962851	0.863454	0.962851	0.938755	0.928715
19^0	0.935743	0.921687	0.944779	0.923695	0.935743	0.958835	0.863454	0.959839	0.934739	0.908635
18^0	0.938755	0.915663	0.940763	0.903614	0.929719	0.956827	0.953815	0.965863	0.929719	0.906627
17^0	0.942771	0.913655	0.935743	0.893574	0.927711	0.953815	0.913654	0.943775	0.928715	0.907631
16^0	0.942771	0.893574	0.932731	0.873493	0.921687	0.954819	0.929718	0.954819	0.939759	0.90261
15^0	0.923695	0.878514	0.934739	0.878514	0.923695	0.951807	0.941767	0.948795	0.938755	0.911647
14^0	0.927711	0.866465	0.929719	0.868473	0.925703	0.949799	0.925702	0.953815	0.925703	0.891566
13^0	0.920683	0.858433	0.927711	0.858433	0.914659	0.945783	0.90261	0.941767	0.923695	0.891566
12^0	0.919679	0.851405	0.917671	0.863453	0.907631	0.940763	0.919618	0.934739	0.914659	0.890562
11^0	0.904618	0.958835	0.914659	0.853413	0.909639	0.947791	0.893574	0.934739	0.925703	0.859438
10^0	0.918675	0.943775	0.913655	0.951807	0.899598	0.947791	0.897536	0.932731	0.911647	0.865462
9^0	0.924699	0.928714	0.915663	0.933734	0.901606	0.940763	0.885542	0.932731	0.88755	0.854418
8^0	0.905622	0.908634	0.908635	0.938755	0.90261	0.932731	0.888554	0.933735	0.898594	0.850402
7^0	0.916667	0.895582	0.90261	0.91465	0.896586	0.935743	0.873493	0.925703	0.883534	0.933734
6^0	0.90261	0.878514	0.88755	0.89257	0.891566	0.933735	0.858433	0.927711	0.896586	0.935742
5^0	0.87249	0.863453	0.883534	0.873493	0.891566	0.923695	0.863453	0.911647	0.873494	0.921686
4^0	0.873494	0.858433	0.874498	0.858433	0.891566	0.919679	0.861445	0.914659	0.85241	0.893574
3^0	0.943775	0.852409	0.943775	0.843373	0.86747	0.918675	0.829317	0.884538	0.861446	0.853413
2^0	0.935742	0.833333	0.935742	0.80522	0.862449	0.911647	0.839357	0.906627	0.963844	0.851405
1^0	0.913654	0.813253	0.934738	0.823293	0.963855	0.907631	0.801204	0.87751	0.952811	0.838353
0^0	0.89257	0.778112	0.928714	0.796184	0.955823	0.883534	0.798192	0.866466	0.93762	0.829317

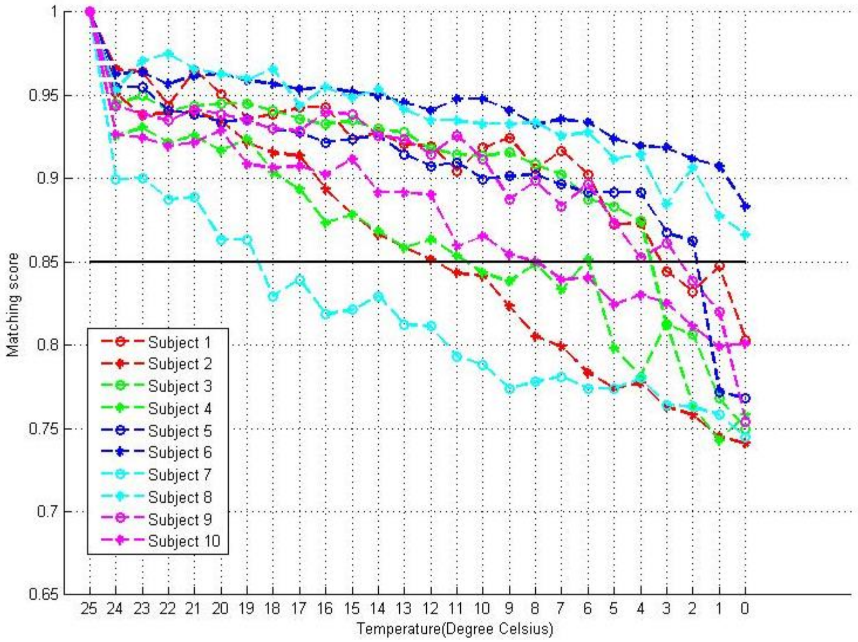


Fig. 5. Graph of Matching Scores Versus Temperature of 10-users/Subjects before Auto-Correction.

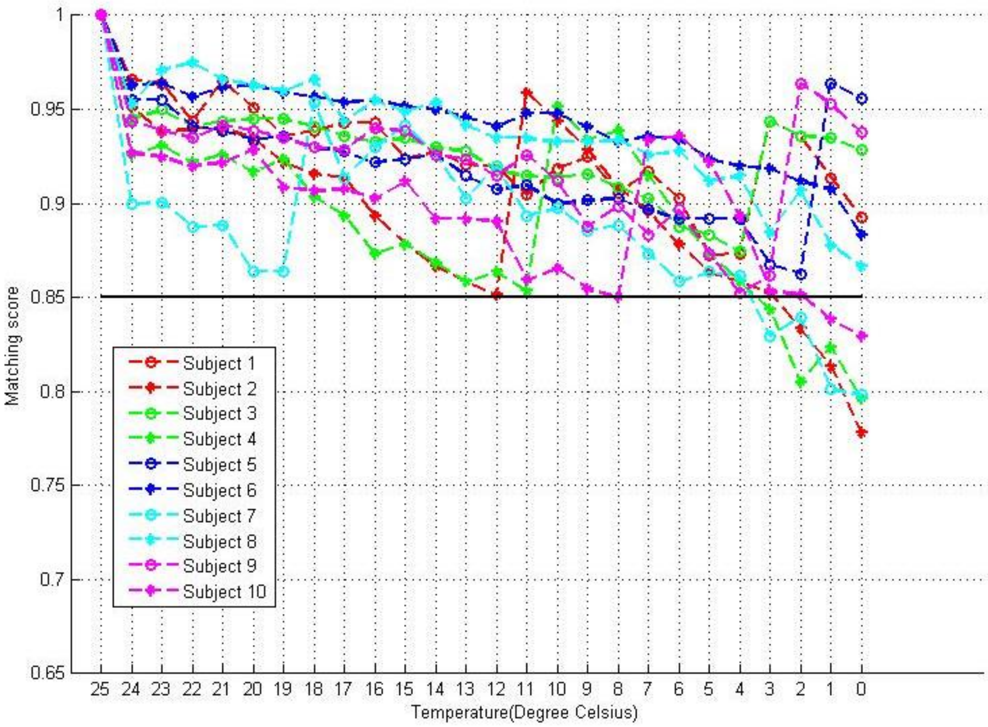


Fig. 6. Graph of Matching Scores Versus Temperature of 10-users/Subjects after Auto-Correction.

TABLE. III. SUMMARY OF SUCCESS AND FAILURE RATE

Total number of users	Without auto-correction				With auto-correction			
	Success		*Failure*		*Success*		*Failure*	
	Number of user	*In percentage (%)*	*Number of user*	*In percentage (%)*	*Number of user*	*In percentage (%)*	*Number of user*	*In percentage (%)*
10	2	20%	8	80%	6	60%	4	40%

TABLE. IV. SUCCESS AND FAILURE RATES ANALYSIS FOR DIFFERENT TEMPERATURE RANGES WITH AND WITHOUT AUTO-CORRECTION

Temperature range	Without auto-correction				With auto-correction			
	Success		*Failure*		*Success*		*Failure*	
	Number of user	*In percentage (%)*	*Number of user*	*In percentage (%)*	*Number of user*	*In percentage (%)*	*Number of user*	*In percentage (%)*
25^0C to 20^0C	10	100%	0	0%	10	100%	0	0%
25^0C to 15^0C	9	90%	1	10%	10	100%	0	0%
25^0C to 10^0C	8	80%	2	20%	10	100%	0	0%
25^0C to 5^0C	6	60%	4	40%	10	100%	0	0%
25^0C to 0^0C	2	20%	8	80%	6	60%	4	40%

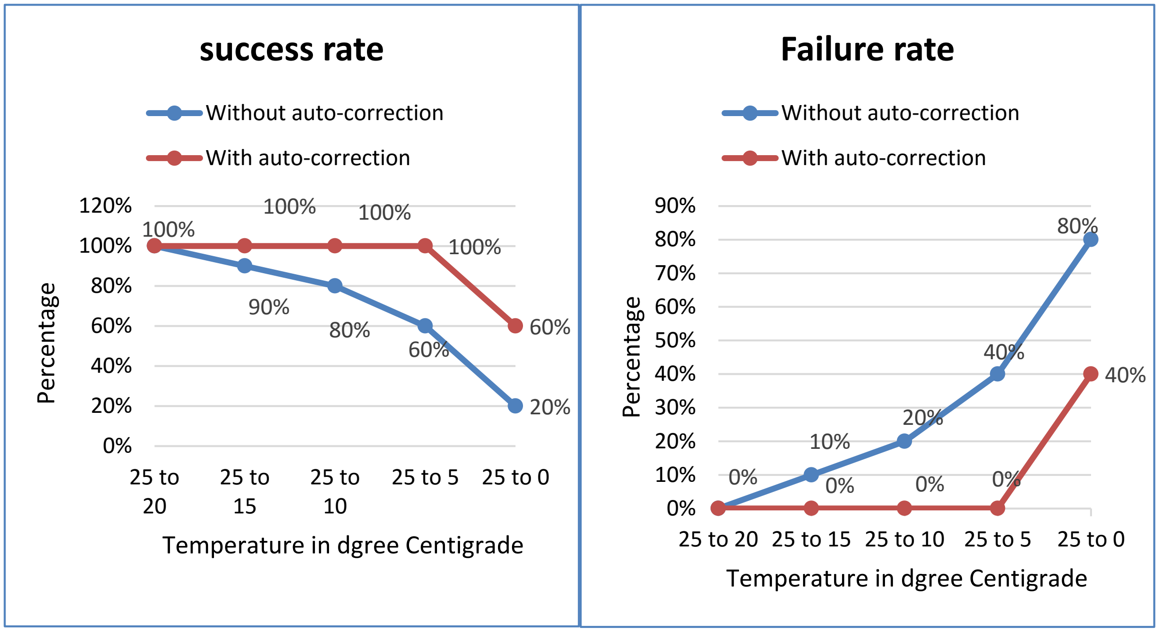


Fig. 7. Graph of Success Rates and Failure Rate with and without Correction between 25^{0}c to 0^{0}c.

IV. Conclusion

The proposed Reference auto-correction algorithm mainly focuses on fingerprint, based time-attendance system on daily basis for user authentication. The principle behind auto-correction algorithm is to correct user reference fingerprint template automatically whenever the rejection of genuine user occurs. Analysis is carried out on 250 fingerprint templates stored in the database of 10-users captured at varying temperature from 25^{0}C to 0^{0}C. Matching score is obtained by comparing the captured fingerprint templates with reference fingerprint template for the respective user without and with using auto-correction algorithm. The experimental result shows 40% improvement in the recognition rate with auto-correction algorithm. Hence, recognition rate of authentication system is improved. The proposed Reference auto-correction technique can also be applied on other Biometric traits like Face recognition, IRIS recognition, palm recognition etc., which are like to vary due aging and due to variation in temperature.

References

[1] Thejaswini P., Srikantaswamy R.S., Manjunatha A.S. (2018),“Impact of Fingerprint Image Quality on Matching Score”, Guru D., Vasudev T., Chethan H., Kumar Y. (eds),Proceedings of International Conference on Cognition and Recognition. Lecture Notes in Networks and Systems, vol 14. Springer, Singapore.

[2] Singh, Jagtar. "A Survey on Fingerprint Recognition Methods." Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org 7.5 (2017).

[3] Yu, Xiaojun, et al. "Contrast Enhanced Subsurface Fingerprint Detection Using High-Speed Optical Coherence Tomography." IEEE Photonics Technology Letters 29.1 (2017): 70-73.

[4] Peralta, Daniel, et al. "Minutiae-based fingerprint matching decomposition: Methodology for big data frameworks." Information Sciences 408 (2017): 198-212.

[5] Li, Hai Ping. "The Application of MATLAB in Automatic Fingerprint Recognition System of Police." MATEC Web of Conferences. Vol. 63. EDP Sciences, 2016.

[6] Yan, Haibin (2016), "Discriminative sparse projections for activity-based person recognition." Neurocomputing, 208, 183-192. doi:10.1016/j.neucom.2015.11.111.

[7] Shanavaz, K. T., and P. Mythili. "A fingerprint-based hybrid gender classification system using genetic algorithm." International Journal of Computational Vision and Robotics 6.4 (2016): 399-413.

[8] Kumar, Sachin, and R. LeelaVelusamy. "Kernel approach for similarity measure in latent fingerprint recognition." Emerging Trends in Electrical Electronics & Sustainable Energy Systems (ICETEESES), International Conference on. IEEE, 2016.

[9] Peralta, Daniel, et al. "A survey on fingerprint minutiae-based local matching for verification and identification: Taxonomy and experimental evaluation." Information Sciences 315 (2015): 67-87.

[10] Sunny Arief SUDIRO, Michel PAINDAVOINE, Tb. Maulana KUSUMA (2007), “Simple Fingerprint Minutiae Extraction Algorithm Using Crossing Number On Valley Structure”, IEEE Workshop on Automatic Identification Advanced Technologies. doi: 10.1109/autoid.2007.380590.

[11] Jing Li, Bao-Liang Lu.” An adaptive image Euclidean distance”, Pattern Recognition, 42(3), 349-357. Doi:10.1016/j.pactog.2008.07.017.

[12] Kouamo, Stephane, and Claude Tangha. "Fingerprint Recognition with Artificial Neural Networks: Application to E-Learning." Journal of Intelligent Learning Systems and Applications 8.02 (2016): 39.

[13] Werner, Gabor Á., and LászlóHanka. "Tuning an artificial neural network to increase the efficiency of a fingerprint matching algorithm." Applied Machine Intelligence and Informatics (SAMI), 2016 IEEE 14th International Symposium on. IEEE, 2016.

[14] Gautam A., Bhateja V., Tiwari A., Satapathy S.C. (2018), “An Improved Mammogram Classification Approach Using Back Propagation Neural Network”, Advances in Intelligent Systems and Computing, vol 542. Springer, Singapore.

[15] M. Mitić, N. Vuković, M. Petrović, Z. Miljković, “Chaotic fruit fly optimization algorithm”, Knowledge-Based Systems (2015), doi: http://dx.doi.org/10.1016/j.knosys.2015.08.010.

Empirical Study of Segment Particle Swarm Optimization and Particle Swarm Optimization Algorithms

Mohammed Adam Kunna Azrag[1], Tuty Asmawaty Abdul Kadir[2]
Faculty of Computer Science & Software Engineering
Universiti Malaysia Pahang, Kuantan, Malaysia

***Abstract*—In this paper, the performance of segment particle swarm optimization (Se-PSO) algorithm was compared with that of original particle swarm optimization (PSO) algorithm. Four different benchmark functions of Sphere, Rosenbrock, Rastrigin, and Griewank with asymmetric initial range settings (upper and lower boundaries values) were selected as the test functions. The experimental results showed that, the Se-PSO algorithm achieved better results in terms of faster convergences in all the testing cases compared to the original PSO algorithm. However, the experimental results further showed the Se-PSO as a promising optimization algorithm method in some other different fields.**

Keywords—Se-PSO; PSO; sphere; Rosenbrock; Rastrigin; Griewank

I. Introduction

Within the last two decades, optimization algorithms with mathematical programing have proved to be effective in solving large complex optimization problems. Recently, swarm intelligence techniques have gained popularity because of their capacity to locate partially optimal solutions for combinatorial optimization problems [1, 2]. These techniques have been applied in various areas, such as economics, engineering, bioinformatics, and industry. These problems are better solved using swarm intelligence techniques because they are usually very hard to solve accurately due to the lack of any precise algorithm to solve them [1, 2]. The swarm intelligence algorithms mainly depend on updating the population of individuals by applying some operators according to the fitness information obtained from the environment. With these updates, the individuals in a population are expected to move towards an optimum solution.

The Particle Swarm optimization (PSO) is one of the popular swarm algorithms which were formulated in 1995 by Dr. Kennedy and Eberhart. The PSO algorithm simulates the flocking behavior of birds and the schooling of fishes in order to achieve complex solutions [3-5]. The PSO algorithm is easy to execute and requires few parameters to be adjusted; it is computationally proficient and has a faster speed and premature convergence towards optima compared to Genetic Algorithm (GA) and Simulating Annealing (SA) algorithm [4]. It also has a flexible and well-balanced mechanism of improving its exploration capabilities [6]. The scenario of PSO is started by initializing a population of random solutions. Each potential solution is assigned with a randomized velocity, and the potential solutions are called particles. Each particle has its own position, velocity, and fitness used to decide its best or bad positions in the solution space. The particle search depends on the personal best position (pbest) and the global best position (gbest) of each particle [7]. Moreover, the ability of PSO to find an optimum solution in reasonable time creates the need for its continuous improvement [8]. However, a segmentation of the PSO is implemented in this study to improve its convergence and accuracy using 4 optimization functions problem.

The rest of this paper is organized thus: the second section discusses the methodology of the PSO and the implemented Se-PSO algorithms. The third section explains the experimental setup and discussion of the results. The fourth section provides the conclusions drawn from the study.

II. Methodology

This paper presents the implementation method of Se-PSO with the comparison of PSO algorithms. The description of Se-PSO and PSO is given in *Pseudo code* in Algo (1 and 2). Se-PSO algorithm is developed according to conventional PSO algorithm. Thus, it has to go through the same procedures initialization of the bird's step, number of birds, iteration, the problem dimension, and the position/velocity updating evaluation processes. However, the segmentation of the problem is added to PSO algorithm in order to reduce the time with a few iterations. The segmentation technique is always providing a fast convergence that able to achieve the best initial local position.

A. PSO Algorithm

The PSO algorithm is simulating the behavior of birds flocking and fish schooling in order to solve optimization problem in D-dimensional search space. This algorithm was proposed in 1995 [9]. The initialization of PSO is start by a group of random particles (solutions) and then searches for the optimum solution by updating generations. Each particle is flown through the solution problem space, having its position based on the information from its own personal best position and the best particle of the swarm. The performance of each particle and how close from the global optimum is calculated using a fitness function of the optimization problem. Each particle i flies through the D-dimensional solution space and maintains the current position x_i of particle i, the personal best position p_i of the particle i, and the current velocity v_i of particle i. The particle keeps track of its coordinates in the solution space which are associated with best solution (fitness) it has achieved so far which stored at each iteration. This value is called pbest (personal best position). Another value of the (best) called gbest (global best position) tracked by the global version of the particle swarm optimizer is the overall best value, and its location obtained so far by any particle in the population. During the particle swarm optimization searching calculation for solution, at each time step, changing the velocity (accelerating) each particle toward its pbest and gbest. Were this calculation mathematically described by this equation:

$$v_i(t+1) = \omega v_i(t) + c_1 r_1 \left(p_i^{best}(t)\right) - c_2 r_2 \left(g_i^{best}(t)\right)$$

$$X_i(t+1 = X_{ir}(t) + v_{ir}(t+1)$$

Where v_i is the velocity of particle i it iteration t, X_i is the position of particle i at iteration t, p_i^{best}= (p_{i1}^{best}, p_{i2}^{best}, …, p_{ir}^{best}) is the personal best position of each particle i till the particle number r and g_i^{best}=(g_{i1}^{best}, g_{i2}^{best}, …, g_{ir}^{best}) is the global best position of the entire particle i till the particle number r ; ω is an inertia weight parameter, $c_1 c_2$ are acceleration coefficients, $r_1 r_2$ are random number between 0 and 1, and r is the dimension in the solution space. The PSO algorithm procedure can be summarized as shown below in Algo 1.

Algo 1

```
1. Start
2. Initialize the $nbird, bird\_s, d, \omega$;
3. Initialize the $v_i, X_i, c_1c_2, r_1r_2$;
4. Calculate the $fitness\ function$
5. If $fitness\ function > 0$
6. For each $Q$;
    a. Iter $Q$=1, $Q++$;
    b. Updating the velocity $v_i$ towards fitness:
    c. $v_i(t+1) = \omega v_i(t) + c_1 r_1 \left(p_i^{best}(t)\right) - c_2 r_2 \left(g_i^{best}(t)\right)$;
    d. Update the position $X_i$ towards fitness:
    e. $X_i(t+1 = X_{ir}(t) + v_{ir}(t+1)$;
7. If $fitness\ function \leq 0$;
    a. Print $g_i^{best}$ of each particles;
8. If $fitness\ function > 0$ return step 2 till the iteration found highly solution or finished.
9. End
```

B. Se-PSO Algorithm

The Segment Particle Swarm algorithm idea is to divide the PSO particles to searching groups that can be considered as segments [10]. The segmentation means to separate the problem into parts or segments which reach the solution easily. In addition, the idea of Segmentation PSO algorithm is to divide the initial values into segments to help PSO particles during the search for the optimal values finding a local best position that may the global best position is around it. The segmentation can be divided into more than two groups based on the dimension.

Considering Fig. 1 is the scenario of parameters segmentation. This scenario contains 3 parameters in search space. Each parameter has its own boundaries. Based on the significance parameter the segment is proposed to find the best position in the parameters boundaries. After the best position is found, the new optimum segment is divided by 2 so the PSO algorithm starts searching from the new segments as described in Fig. 2.

Each group of particles considered a segment while the procedures for finding the optimal solution (optimal segments) is following PSO algorithm then the optimal segment for the initial parameters will be used as the new initial parameters later PSO search in that range toward the optimal solution. The segmentation can be described in Fig. 3.

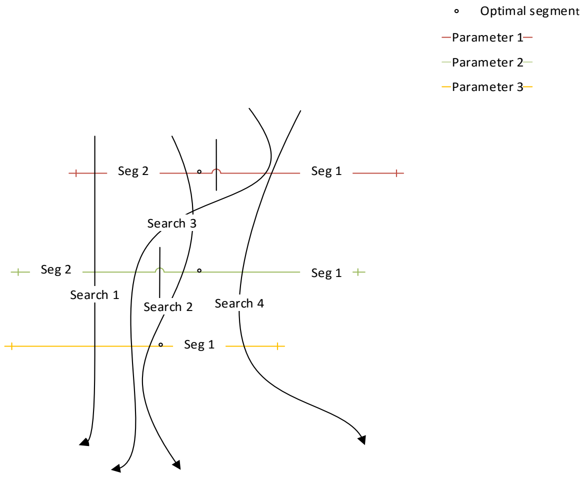


Fig. 1. Parameter Segment.

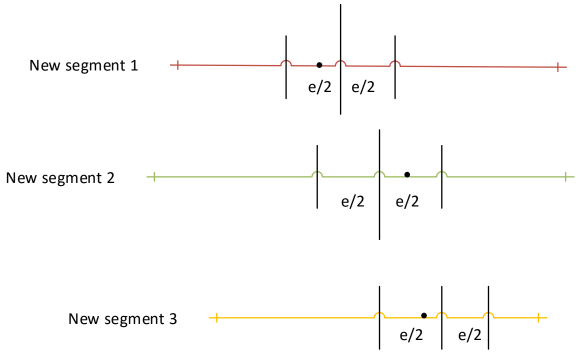


Fig. 2. New Segment.

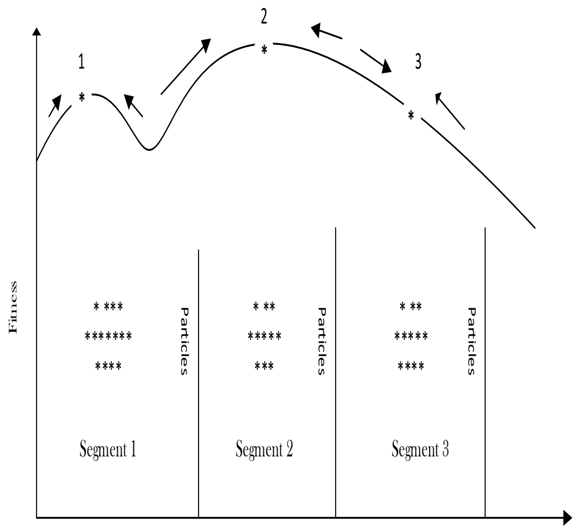


Fig. 3. Segment Particles.

In the above Fig. 3, the segments easily locate many local points where point 2 is the best local points, as it's the best individual position of the particle and the best position of the entire swarm should be modified to achieve the best optimal segment as a fellow in Eq. (1).

$$Seg\ length = \frac{initial\ limits}{nu\ of\ segments} \quad (1)$$

Then eq. (3) should be modified according to the segmentation changes and described as follow in Eq. (2) and (3).

$$v_i\ (t+1,j) = \omega v_i\ (t,j) + c_1\, r_1\ (p_i\ (t,j) - X_i\ (t,j)) + c_2\, r_2\ (G_i\ (t,j) - X_i\ (t,j)) \quad (2)$$

$$X_i(t+1,j) = X_i(t,j) + v_i(t+1,j) \quad (3)$$

Hence the optimal segment can be described in Eq. (4) below:

$$optimal\ segment = \frac{optimal\ X_{ij} \mp segemnt\ length}{2} \quad (4)$$

Where j is the number of segments.

Algo 2

```
1   BEGIN
2    Initialize $S, D, \omega$;
3    initialize $v_i$, $X_i$, $c_1$, $c_2$, $r_1$, $r_2$, number_segment;
4     Segment length=initial value/number_segment;
5   Adopt the $b_{e1}$:$b_{en}$ parameters boundaries;
6       For j= 1 to number of segment;
7       Determine initial $fitness$ for segment $J$;
8       Assume $Best\ fitness = initial\ fitness$;
9           If $fitness\ >$ Best $fitness$;
10          For each $S$;
11              iter $S$=1, $S$++;
12              Updating the velocity $V_i$ toward fitness:
        $v_i\ (t+1,j) = \omega v_i\ (t,j) + c_1\, r_1\ (p_i\ (t,j) - X_i\ (t,j)) + c_2\, r_2\ (G_i\ (t,j) X_i\ (t,j))$;
13          Update the position $X_{i,}j$ toward fitness:
            $X_i(t+1,j) = X_i(t,j) + v_i(t+1,j)$;
14          End if,
15        If fitness $\leq$ Best $fitness$;
              Print $G_i best$ of each particles;
              Update the $b_{e1}$:$b_{en}$based of $G_i best$ of each kinetic parameters;
16            If fitness $>$ Best $fitness$ return step 2 till
          the iteration is finished or discover high-quality
          solution;
17         End if,
18         End if,
19         Global point(j)= $G_i best$;
20         Next, j;
21         Optimal_segment=max (Global point) ±
          segment length/2;
22         Repeat algorithm 1 for the new initial values;
23        End.
```

III. Experimental Setup

For comparison, for nonlinear functions are used here. The first function is the Sphere function described by equation ($f(x)$):

$$f(x) = \sum_{s=1}^{n} X^2$$

Where $X = (X_1, X_2, \dots, X_n)$ is an n-dimensional real-valued vector. The second function is the Rosenbrock function described by equation ($f_1(x)$):

$$f_1(x) = \sum_{i=1}^{n} (100(X_i - X_i^2)^2 + (X_i - 1)^2)$$

The third function is generalized Rastrigrin function described by equation ($f_2(X)$):

$$f_2(X) = \sum_{i=1}^{n} (X_i^2 - 10\cos(2\pi X_i) + 10)$$

The last function is the generalized Griewank function described by equation ($f_3(X)$):

$$f_3(X) = \frac{1}{4000} \sum\nolimits_{i=1}^{n} X_i^2 - \prod_{i=1}^{n} \cos\left(\frac{X_i}{\sqrt{i}}\right) + 1$$

TABLE I. LOWER AND UPPER BOUNDARIES

Function	Lower and upper values
$\boldsymbol{f}$	[-5, 5]
$\boldsymbol{f_1}$	[-5, 10]
$\boldsymbol{f_2}$	[-5.12, 5.12]
$\boldsymbol{f_3}$	[-600, 600]

Following the suggestion in [11] and for the purpose of comparison, the lower and upper values are selected to be as the original values. Table I lists the initialization of the upper and lower values of the four functions.

As in the above Table I, for each function, the maximum number of iteration is set to 50, 100, 150 in the both algorithms. The number of birds is set to 20, 40, 60, and 80. In order to compare the both algorithms inertia weight is adapted to 0.9. The learning factors $c_1\, c_2 \approx 4$. The dimension is 10 for both. Each algorithm was tested 10 times to get the Mean global best position.

IV. RESULTS AND DISCUSSION

In Sphere function, the global best position of Se-PSO showed a very improved result in short time when compared to PSO. Moreover, as shown in Tables II and III the convergence speed of the Se-PSO towards the optimum values was faster when compared to PSO. It's however, noticeable that the convergence of Se-PSO was quick in the all function but slowed down searching large space for the global best position before to be decided by Se-PSO algorithm as it approaches the optimal. The Se-PSO it took 0.213s to reach the best global position but the PSO reached the global best position in 0.033s. Where, Se-PSO takes 0.004s to decide the global best position while PSO takes 0.013s are depicted in Tables II and III in the self –time column. Moreover, Se-PSO searching large space almost 7 times of PSO as it described in the Calls column (4920 and 650) respectively.

Looking at Table IV where Sphere function was tested using Se-PSO and PSO on a 10-dimensional space with 5, 15, and 25 bird-step and 5, 15, and 25 iterations, the global best position of Se-PSO has very good result that is almost 4 times as good as the result of PSO. Moreover, the convergence speed of Se-PSO toward the optimum values are faster than those of PSO as showed in Tables II and III. Moreover, it is easy to observe that Se-PSO convergences quickly but slows its convergence speed down when reaching the optimum. It takes 0.004s to get the best global position while PSO takes 0.013s as can be seen clearly in Tables II and III. Finally, the total performance speed of the both algorithm is 0.213s with 25 iterations for Se-PSO and 0.033s only for PSO.

In Rastrigin function, the global best position of Se-PSO showed a very improved result in short time when compared to PSO in Table VII. Moreover, as shown in Tables V and VI the convergence speed of the Se-PSO towards the optimum values was faster when compared to PSO. It's however, noticeable that the convergence of Se-PSO was quick in the all function but slowed down searching large space for the global best position before to be decided by Se-PSO algorithm as it approaches the optimal. The Se-PSO it took 0.363s to reach the best global position but the PSO reached the global best position in 0.048s. Where, Se-PSO takes 0.001s to decide the global best position while PSO takes 0.013s are depicted in Tables V and VI in the self –time column. Moreover, Se-PSO searching large space almost 7 times of PSO as it described in the Calls column (4920 and 650), respectively.

TABLE II. SE-PSO CONSUMPTION FOR SPHERE FUNCTION

Function	Calls	Total time	Self-time
Se-PSO	1	0.213s	0.004s
PSO	3	0.209s	0.030s
Sphere	4920	0.179s	0.179s

TABLE III. PSO CONSUMPTION FOR SPHERE FUNCTION

Function	Calls	Total time	Self-time
PSO	1	0.033s	0.013s
Sphere	650	0.020s	0.020s

TABLE IV. SE-PSO AND PSO RESULTS FOR SPHERE FUNCTION

Bird step	Dimension	Iteration	$\boldsymbol{g^{Best}}$ of Se-PSO	$\boldsymbol{g^{Best}}$ of PSO
5	10	5	1.01335e-5	0.0701
		15	2.12225e-06	0.0517
		25	4.25101e-7	0.0074
15	10	5	3.25661e-07	0.0158
		15	4.25861e-08	0.0132
		25	1.05843e-08	0.0013
25	10	5	1.00035e-08	0.0001
		15	2.03589e-09	0.0011
		25	1.23565e-0.9	0.00002

TABLE V. SE-PSO CONSUMPTION FOR RASTRIGIN FUNCTION

Function	Calls	Total time	Self-time
Se-PSO	1	0.363s	0.001s
PSO	3	0.362s	0.028s
Rastrigin function	4920	0.334s	0.334s

TABLE VI. PSO CONSUMPTION FOR RASTRIGIN FUNCTION

Function	Calls	Total time	Self-time
PSO	1	0.048s	0.013s
Rastrigin Function	650	0.035s	0.035s

TABLE VII. SE-PSO AND PSO RESULTS FOR RASTRIGIN FUNCTION

Bird step	Dimension	Iteration	$\boldsymbol{g^{Best}}$ of Se-PSO	$\boldsymbol{g^{Best}}$ of PSO
5	10	5	1.1237e-05	1.0835
		15	8.3919e-06	0.2655
		25	3.5864e-06	0.9891
15	10	5	1.6515e-06	0.0028
		15	2.2586e-07	0.0359
		25	4.2056e-07	0.0026
25	10	5	3.0125e-08	0.0085
		15	2.0123e-08	0.0064
		25	1.0213e-09	0.003

In the Rosenbrock function, the global best position of Se-PSO showed a very improved result in short time when compared to PSO in Table X. Moreover, as shown in Tables VIII and IX, the convergence speed of the Se-PSO towards the optimum values was faster when compared to PSO. It's however, noticeable that the convergence of Se-PSO was quick in the all function but slowed down searching large space for the global best position before to be decided by Se-PSO algorithm as it approaches the optimal. The Se-PSO it took 0.158s to reach the best global position but the PSO reached the global best position in 0.03s. Where, Se-PSO takes 0.004s to decide the global best position while PSO takes 0.013s are depicted in Tables VIII and IX in the self –time column. Moreover, Se-PSO searching large space almost 7 times of PSO as it described in the Calls column (4920 and 650) respectively.

Looking at Table X above, where Rosenbrock function was tested using Se-PSO and PSO on a 10-dimensional space with 5, 15, and 25 bird-step and 5, 15, and 25 iterations, the global best position of Se-PSO has very good result that is almost 4 times as good as the result of PSO. Moreover, the convergence speed of Se-PSO toward the optimum values are faster than those of PSO as showed in Tables VIII and IX. Moreover, it is easy to observe that Se-PSO convergences quickly but slows its convergence speed down when reaching the optimum. It takes 0.004s to get the best global position while PSO takes 0.013s as can be seen clearly Tables VIII and IX. Finally, the total performance speed of the both algorithm is 0.158s with 25 iterations for Se-PSO and 0.030s only for PSO.

The global best position of Se-PSO showed a very improved result in short time when compared to PSO in Griewank function depicted in Table XIII. Moreover, as shown in Tables XI and XII the convergence speed of the Se-PSO towards the optimum values was faster when compared to PSO. It's however, noticeable that the convergence of Se-PSO was quick in the all function but slowed down searching large space for the global best position before to be decided by Se-PSO algorithm as it approaches the optimal. The Se-PSO it took 0.037s to reach the best global position but the PSO reached the global best position in 0.014s. Where, Se-PSO takes 0.002s to decide the global best position while PSO takes 0.013s are depicted in Tables XI and XII in the self –time column. Moreover, Se-PSO searching large space almost 7 times of PSO as it described in the Calls column (4920 and 650), respectively.

TABLE VIII. SE-PSO CONSUMPTION FOR ROSENBROCK FUNCTION

Function	Calls	Total time	Self-time
Se-PSO	1	0.158s	0.004s
PSO	3	0.154s	0.026s
Rosenbrock function	4920	0.128s	0.128s

TABLE IX. PSO CONSUMPTION FOR ROSENBROCK FUNCTION

Function	Calls	Total time	Self-time
PSO	1	0.030s	0.013s
Rosenbrock function	650	0.016s	0.016

TABLE X. SE-PSO AND PSO RESULTS FOR ROSENBROCK FUNCTION

Bird step	Dimension	Iteration	g^{Best} of Se-PSO	g^{Best} of PSO
5	10	5	0	0.8743
		15	0	0.3272
		25	0	0.0545
15	10	5	0	0.6533
		15	0	0.942
		25	0	0.7522
25	10	5	0	0.539
		15	0	0.6945
		25	0	0.2477

TABLE XI. SE-PSO CONSUMPTION FOR GRIEWANK FUNCTION

Function	Calls	Total time	Self-time
Se-PSO	1	0.037s	0.002s
PSO	3	0.035s	0.022s
Griewank function	4920	0.013s	0.013s

TABLE XII. PSO CONSUMPTION FOR GRIEWANK FUNCTION

Function	Calls	Total time	Self-time
PSO	1	0.014s	0.013s
Griewank function	650	0.001s	0.001

TABLE XIII. SE-PSO AND PSO RESULTS FOR GRIEWANK FUNCTION

Bird step	Dimension	Iteration	g^{Best} of Se-PSO	g^{Best} of PSO
5	10	5	4.5368e-05	0.1176
		15	8.1222e-07	0.0446
		25	4.1117e-09	0.0023
15	10	5	3.6206e-09	0.0075
		15	8.4574e-10	0.3777
		25	5.2106e-11	0.3637
25	10	5	9.2952e-12	0.0137
		15	3.5258e-13	0.0024
		25	4.2586e-14	0.0019

V. CONCLUSION

The Se-PSO algorithm was introduced in this study through the incorporation of a segmentation solution during the searching process by the particles into the original version of the PSO. After this, the best segment position of Se-PSO algorithm was projected as the new search dimension to find the optimum values. To investigate the proposed method, four functions were employed. The results of the experiments showed that the Se-PSO exhibited fast convergence when compared with the ABO. Although it was observed that the Se-PSO algorithm worked efficiently, the need for further experiments on more complex multimodal, separable and non-separable functions cannot be ruled out in order to further validate the search capacity of the Se-PSO. Moreover, there is a need to apply the proposed Se-PSO to other optimization search landscapes such as urban transportation problems, job scheduling, dynamic modeling, and vehicle routing.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from the Universiti Malaysia Pahang and FSKKP under PGRS190312 and FRGS under Grant RDU 160101. The authors would like to thank Dr. Tuty Asmawaty Abdul Kadir, who supervised this work.

REFERENCES

[1] Soliman, Soliman Abdel-Hady, and Abdel-Aal Hassan Mantawy. Modern optimization techniques with applications in electric power systems. Springer Science & Business Media, 2011.

[2] Venter, Gerhard. "Review of optimization techniques." Encyclopedia of aerospace engineering (2010).

[3] Kennedy, James. "Particle swarm optimization." Encyclopedia of machine learning. Springer, Boston, MA, 2011. 760-766.

[4] Eberhart, Russell C., and Yuhui Shi. "Comparison between genetic algorithms and particle swarm optimization." International conference on evolutionary programming. Springer, Berlin, Heidelberg, 1998.

[5] Khare, Anula, and Saroj Rangnekar. "A review of particle swarm optimization and its applications in solar photovoltaic system." Applied Soft Computing 13.5 (2013): 2997-3006.

[6] Banks, Alec, Jonathan Vincent, and Chukwudi Anyakoha. "A review of particle swarm optimization. Part I: background and development." Natural Computing 6.4 (2007): 467-484.

[7] Banks, Alec, Jonathan Vincent, and Chukwudi Anyakoha. "A review of particle swarm optimization. Part II: hybridisation, combinatorial, multicriteria and constrained optimization, and indicative applications." Natural Computing 7.1 (2008): 109-124.

[8] Shi, Yuhui. "Particle swarm optimization: developments, applications and resources." evolutionary computation, 2001. Proceedings of the 2001 Congress on. Vol. 1. IEEE, 2001.

[9] Poli, Riccardo, James Kennedy, and Tim Blackwell. "Particle swarm optimization." Swarm intelligence 1.1 (2007): 33-57.

[10] Jaber, Aqeel S., Abu Zaharin Ahmad, and Ahmed N. Abdalla. "A new parameters identification of single area power system based LFC using Segmentation Particle Swarm Optimization (SePSO) algorithm." Power and Energy Engineering Conference (APPEEC), 2013 IEEE PES Asia-Pacific. IEEE, 2013.

[11] Shi, Yuhui, and Russell C. Eberhart. "Empirical study of particle swarm optimization." Evolutionary computation, 1999. CEC 99. Proceedings of the 1999 congress on. Vol. 3. IEEE, 1999.

An Efficient Deep Learning Model for Olive Diseases Detection

Madallah Alruwaili[1]
Computer Engineering and Networks Department
Jouf University
Sakaka, KSA

Sameh Abd El-Ghany[3]
Department of Information Systems
Mansoura University, Mansoura 35516, Egypt
Information Systems Department
Jouf University, Sakaka, KSA

Saad Alanazi[2]
Computer Science Department
Jouf University
Sakaka, KSA

Abdulaziz Shehab[4]
Department of Information Systems
Mansoura University, Mansoura 35516, Egypt
Department of Computer Science
College of Science and Arts, Jouf University, KSA

***Abstract*—Worldwide, plant diseases adversely influence both the quality and quantity of crop production. Thus, the early detection of such diseases proves efficient in enhancing the crop quality and reducing the production loss. However, the detection of plant diseases either via the farmers' naked eyes or their traditional tools or even within laboratories is still an error prone and time consuming process. The current paper presents a Deep Learning (DL) model with a view to developing an efficient detector of olive diseases. The proposed model is distinguishable from others in a number of novelties. It utilizes an efficient parameterized transfer learning model, a smart data augmentation with balanced number of images in every category, and it functions in more complex environments with enlarged and enhanced dataset. In contrast to the lately developed state-of-art methods, the results show that our proposed method achieves higher measurements in terms of accuracy, precision, recall, and F$_1$-Measure.**

Keywords—Deep learning; AlexNet; conventional neural networks; plant diseases; olive; feature extraction

I. Introduction

Nowadays and with the advancement of technology such as digital cameras and other new portable devices in image processing, a growing interest has emerged to construct methods that enhance crop production on both quantitative and qualitative bases. A large number of plant diseases contribute to the reduction of the crop production. These diseases mainly influence the state and color of plant leaves, roots, buddings, flowers and fruits. Due to the similar patterns of diseases, it has been proven difficult to point out these minimal differences, rendering their visibility a major challenge. Likewise, inexperienced farmers find it difficult to detect plant diseases with their naked eye. If they were successful, they may not be aware of the appropriate treatment. Thus, the early diagnosis and treatment of these diseases can minimize the losses in the whole crop. To this end, new trends of research have considered the validity of automated methods in detecting plant diseases. The most common techniques for disease detection are machine learning coupled with image processing. These techniques monitor, measure and analyze various images exhibiting common plant diseases. As shown in "Fig. 1", they involve steps that should be taken into consideration for the early identification and diagnosis of plant diseases.

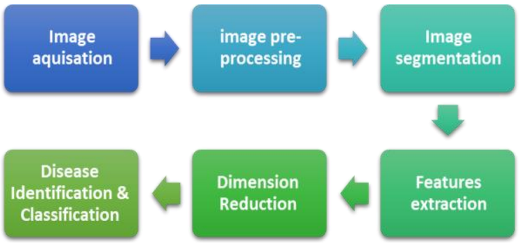


Fig. 1. Basic Steps of Plant Disease Identification and Classification.

Image acquisition aims to obtain and collect images that help in system training. Thus, it is a crucial step given that the accuracy of the system heavily relies on the samples of image for the training purposes [1]. Current studies use either self-collected image datasets or benchmark datasets such as IPM Images, PlantVillageImages, and APS Image database. (2) As for Image Pre-processing, it seeks to improve the accuracy of disease typology. It includes noise removal, image enhancement techniques, Image quantization, spatial filtering, background removal, resizing and cropping operations. (3) Image Segmentation approach is the third step that can be generally split into two categories: (i) edge detection or (ii) pixel classification. Histogram thresholding, edge detection, Region of Interest (ROI), Otsu's, K-means and Fuzzy c-means are examples for segmentation methods. (4) Feature extraction: in order to identify leaf diseases, we should appeal to appropriate features as a distinguishing descriptor of disease typology. Concerning feature extraction step, it is among the most helpful steps in disease detection and classification as it plays a key role in distinguishing one disease from another. An inappropriate or excessive use of features may cause the classification to be over fitting and require long search time. Thus, it is highly recommended to resort to an effective descriptor of the various diseases. The image features can be subsumed under three categories: color, texture, and shape. (5) Dimension Reduction: As a prior step, it is better to pre-process the training data by reducing the dimension of the feature vector so as to maximize the speed and efficiently of the search. Correlation feature selection (CFS) is a central problem to identify the appropriate features for developing a classification model for a particular task via a correlation based method. A good feature subset can enhance the model interpretability, save training time and improve the generalizations by minimizing overfitting. The core assumption is that a good feature set is closely correlated with the class and uncorrelated with other classes. (6) Disease Identification & classification: a number of automatic methods/models for the detection and classification of plant diseases have been put forward in the recent decades to overcome the limitation of human based visual detection. These methods are based on different machine learning algorithms such as neural network [2], Fuzzy logic, K- nearest neighbor, support vector machine [3], AdaBoost [4], rule base [5], and deep learning [6].

In this paper, we propose an enhanced Convolutional Neural Networks (CNNs) named AlexNet for olive disease detection and classification. Its main contribution is the improvement of the accuracy of olive diseases diagnosis. According to FAO, Olive trees are among the most cultivated plants on the globe and the number of olive trees planted in 2014 is estimated at about 10.2 million hectares. However, olive trees are currently under the threat of a variety of diseases that affect its growth and quality such as canker, Anthracnose, Peacock spot, etc. compared to state-of-art methods, with the aid of an efficient parameterized transfer learning model and a smart data augmentation technique, the results show a predomination although our proposed model functions in more complex environments. The organization of the paper proceeds as follow. Section II provides the relevant literature. An overview of the system framework is presented in Section III. The proposed technique will be laid out in Section IV. In Section V, we will provide a detailed description of the experimentation processes. The results will be discussed in Section VI. Concluding remarks and future work will be given in Section VII.

II. Related Work

In response to the costly losses in agricultural production caused by insects and plant diseases, new technological methods have been designed. Computer science is among the disciplines that address these concerns and provide solutions to them. Machine learning, for an instance, plays a key role in detecting such pests and epidemics. In the past decades, a considerable volume of studies with different machine learning algorithm have been executed for Plant disease detection under different environmental conditions, in different countries, and for different plants such as tomato [7], potato [8], rice [9], cassava [10], mango [11], apple [12, 13] , general plants [14, 15], and Olive [16, 17], etc. Jagan Mohan et al.[4] presented a system that firstly used SIFT to extract featured from the paddy plant; secondly the AdaBoost classifier was used for disease detection with identification rate 83.33%. After identification, the diseases are recognized with the use of SVM classifier with a recognition rate 91.10% for SVM and 93.33% for K-NN. A system that employ fuzzy logic to detect scab disease in apple is presented in [18]. In this study Blob analysis method was used to for feature extraction and the system get accuracy 91.66% for disease classification. Different deep learning architectures were used in [7] for identifying tomato stresses. The author concluded that Faster RCNN with VGG-16 performed better than other architectures used in the experiments. Monzurul Islam, et al. [8], developed a system to classify the potato diseases . The system firstly mask out the background as well as the green region of the leaves to extract the region that contains disease symptoms. After that two feature extraction methods are used: color and texture, then multiclass SVM classifier is employed for image classification. The accuracy of the system is 95%. In [9], AlexNet deep learning is used to classify rice plant image to three classes, (a) it is infested with golden apple snails; (b) it is afflicted with diseases; and (c) it is normal and healthy; the system provided 91.23% accuracy. Ramcharan, et al. [10] trained a Tensor flow model of CNN to detect and identify foliar symptoms of diseases in cassava. Thereafter, the trained model was employed in a mobile app. The CNN detection model achieves 94 ± 5.7%. Singh, et al. [11] presented a multilayer convolutional neural network (MCNN) based approach for identifying Anthracnose fungal disease affect Mango leaves, the system get average accuracy of 97.13%. Detection of anthracnose lesion in apple fruit using adapted DenseNet model was presented in [12], and it achieved an overall accuracy of 95.57% for disease identification. Apple leaf diseases using deep-CNNs is proposed in [13], in this system; GoogLeNet Inception structure and Rainbow concatenation (VGG-INCEP model.) are employed to detect five apple leaf diseases (Alternaria leaf spot, Brown spot, Mosaic, Grey spot, and Rust) The classifier achieved a recognition accuracy of 97.14%. AlexNet and GoogLeNet are used in [14] for classification of 54,306 images from PlantVillage to healthy and infected plant leaves. Their model achieved an accuracy of 99.35%. However, their proposal has dramatically collapsed when tested on images taken under different condition. Sladojevic, et al. [15] developed a model using CaffeNet for recognizing 14 different types of plant diseases from healthy leaves in Peach, Apple, and Grapevine. They achieved an average of 96.3% accuracy in their experimental analysis. Al-Tarawneh, Mokhled [14] worked on olive leaves spot diseases and proposed a novel technique which is a combination of auto-cropping segmentation and fuzzy c-means clustering. The segmentation part is done by Automatic polygon cropping ROI, while Fuzzy C-means clustering classifier was used to classify the diseases by comparing c- means clustering with k-means clustering with performance parameters like speed and accuracy. The results show that Fuzzy c-means clustering was found superior than the K-means clustering. A DL model for identification of olive ‘quick decline’ syndrome called ‘abstraction-level fusion’ was proposed in [17]. The rate of detection was over 98%, despite the presence of disease in olive fruits. Yet, this system identifies the diseases on the olive leaves only.

III. System Framework

The present paper aims for a smart system capable of processing olive leaf images with the support of machine learning algorithms and detecting different symptoms of olive diseases. A novel deep learning framework based on AlexNet model was provided to organize the learning process using the transfer learning approach. The proposed model, discussed in more details in the next section, discovers low-level features plant images and in turn detects the disease. The system overview is described as follows: (1) for this study, the data is taken from plant-village dataset. Subsequently, it is enhanced with the olive data collected for Aljouf laboratory. (2) Leaf images are pre-processed via a small window median filter. The filtering process removes noise, and then the images are resized to 256 pixels × 256 pixels. (3) Furthermore, the images are processed with the proposed AlexNet Model. (4) Diagnosis and feedback. "Fig. 2" shows the working steps of the system overview.

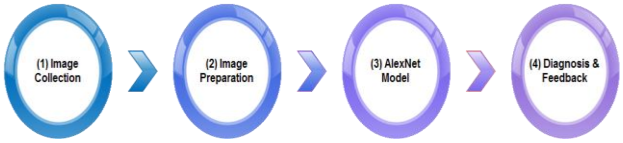


Fig. 2. System Overview.

AlexNet model starts from data collection until up to disease detection process. Occasionally, one of the key steps seeks to efficiently prepare the collected data in a manner fit for the model and assist it in obtaining accurate results. Data preparation includes data augmentation, equalizing the number of images in each class, simple filtering and de-noising. Afterwards, the dataset is divided into training set (80%) and testing set (20%). Optimizing the DL network parameters is accomplished through the training process. The more learned, the more accurate the model becomes in mapping input into its desired output. The trained DL model is then tested against the remaining 20% unseen images. As a final step, after carrying out a number of experiments and monitoring the output results, the DL network is said to be converged and deployed.

IV. Proposed Model

A. AlexNet Description

Initially, at low-level, a convolutional neural network (CNN) is fed by image's pixel representation defined in PlantVillage dataset. Layers are interconnected together in a Multi-Layer (ML) architecture. The network is in charge of converting the input visual stimulus into non-local signals. Gradually, the abstraction-level of the signal becomes more complex due to passing through succeeding layers. Low-complex features like edges, corners, intensity values, and texture are captured by initial layers while more complex features are formed in a step-by-step format at the higher layers of abstraction. Deep learning (DL) is based on ML techniques where current succeeding layer depends on the output reached from the previous layer. In other words, every two subsequent layers are connected together by neurons. These neurons are nonlinearly transformed layer features. Parameters like weights and biases should be carefully assigned as it determines the accuracy of classification. DL networks are bi-directional neural networks. In the forward direction, image's pixel representation is passed to the input layer which in turn transforms it to next layer. The transformation continues through hidden layers until the output layer. At the output layer, also the decision layer, a decision is made and the residual error is calculated between desired output and the output actually obtained. In the backward direction, the error is fed and backed in order to correct the parameters reversely from output layer to input layer. The model keeps moving forward and backward until achieving satisfied results. Typically, a learning algorithm called back-propagation is utilized for such problems. Stochastic gradient descent (SGD) and its variants, such as mini batch gradient descent [19], ADAM [20], and ADMM [21], have been used to train DNNs. The AlexNet architecture [22] follows the same design pattern as the LeNet-5 [23] architecture. It is a set of stacked convolution layers followed by one or more fully connected layers as shown in "Fig. 3". Optionally, the convolution layers may have normalization and pooling layers. Thereafter, Rectified Linear Unit (ReLu) non-linear activation function is usually used for all the network layers.

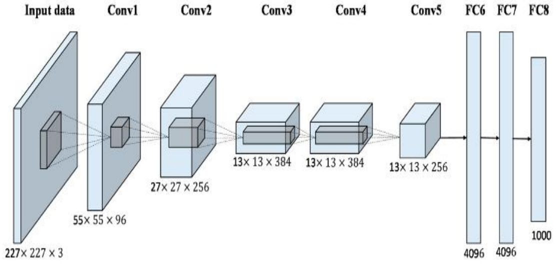


Fig. 3. Classic Structure of AlexNet Deep Neural Network [24].

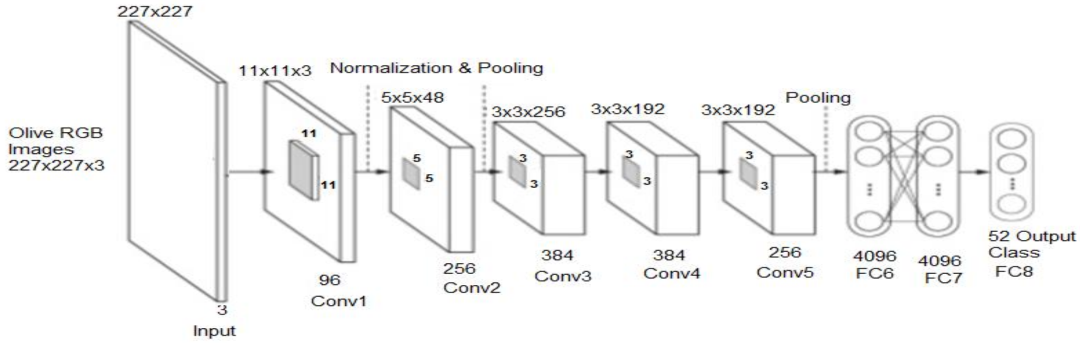


Fig. 4. Structure of the Convolutional Neural Network Employed in this Work.

As demonstrated in "Fig. 4", our adapted version AlexNet consists of 5 convolution layers (conv1—5), followed by 3 fully connected layers (fc6, fc7, and fc8), and a softMax layer comes at the end. The first two convolution layers (conv1 and conv2) are each followed by both cross channel normalization layer (5 channels per element) and 3x3 max pooling layer. Meanwhile, the final convolution layer (conv5) is followed by only one pooling layer. The final fully connected layer (fc8) has 52 outputs which equal the total number of classes in our enhanced dataset. Dependently, normalization of fc8 output is exponentially done at the softMax layer. ReLu proceeds for 7 layers of AlexNet from Conv1 through FC7. To reduce the overfitting, both fc6 and fc7 have a dropout ratio of 0.5 overfitting occurs when the model learns massive details about the training data. An overfitting overcome method known as 'sgdm' is used. In addition, novel regularization techniques such as dropout [25] have emerged to reduce overfitting. To build a robust and unbiased model that discovers true parameters, a huge amount of input data with a high changeability must be guaranteed. However, it is an arduous task to obtain a dataset that could cover such changeability in most common realistic applications. To address this issue, as we have not a sufficient large olive dataset, we have to augment the dataset. Data augmentation is accomplished to aid the proposed DL model reaching higher results compared to other methods. Unlike [15], at our Model's training stage, classes suffering a small number of images are geometrically modified with a random transformation such as changing the intensity of RGB channels, flipping, translations or rotations. Moreover, each class has the same number of training images with equal percentage of disease classes.

B. *Artificial Data Augmentation*

Some DL models may not be sustainable to natural disorders such as illumination, perspective, and position variability in test images. Ideally, good DL models should overcome both underfitting and overfitting problems. Yet, it is an overambitious goal to achieve in practice. It requires successive trials to reach such a goal. The issue of overfitting occurs when the model learns massive details about the training data. An overfitting overcome method known as 'sgdm' is used. In addition, novel regularization techniques such as dropout [25] have emerged to reduce overfitting. To build a robust and unbiased model that discovers true parameters, a huge amount of input data with a high changeability must be guaranteed. However, it is an arduous task to obtain a dataset that could cover such changeability in most common realistic applications. To address this issue, as we have not a sufficient large olive dataset, we have to augment the dataset. Data augmentation is accomplished to aid the proposed DL model reaching higher results compared to other methods. Unlike [15], at our Model's training stage, classes suffering a small number of images are geometrically modified with a random transformation such as changing the intensity of RGB channels, flipping, translations or rotations. Moreover, each class has the same number of training images with equal percentage of disease classes.

V. Experimentation Details

A. *Dataset Description*

Original PlantVillage dataset has 54,306 images of plant leaves distributed over different 38 classes of 14 crop species and 26 diseases. During the implementation of our experiments, the model is deployed over PlantVillage dataset without any interventions. Nonetheless, due to the absence of olive diseases in it, it was inevitable to enhance the dataset to be further considered. Extra 14 classes of different olive's images are added to PlantVillage dataset and hence 52 classes are formed. Olive Plants within groups were selected from the same olive tree age (25–30 years). To evaluate the presence of a specific disease like Anth, Aspid, Cankers, Frost, Gloe, Hail, etc. in olive trees, sampling is executed according to symptoms in January 2018. "Fig. 5" shows a sample of olive leaf diseases.

B. *Performance Metrics*

To evaluate the proposed model, equations (1)-(4) are employed. All of the following metrics are exposed as percentages.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+ FN}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{F}_1 = 2 * \frac{\text{Preceision}*\text{Recall}}{\text{Preceision+Recall}} \quad (4)$$

C. *Implementation Details and Parameters*

Due to a controlled collection of PlantVillage dataset, it is expected to obtain highly accurate results. Therefore, we are motivated to further assess the model's performance on an enhanced version of PlantVillage dataset which contains olive images. Such images are scarce (verified 14 class of 120 images). In order to record the olive results, the network is re-trained on the new updated version of the dataset. To train the model, parameters summarized in "Table I" are utilized. With the purpose of verification, we perform 10 trials with random sets of 80% for training data and 20% for testing data.

TABLE. I. AlexNet Parameters

Factor	**Value**
'WeightLearnRateFactor'	20
'BiasLearnRateFactor'	20
'InitialLearnRate'	1e-4
'MiniBatchSize'	10
'MaxEpochs'	6
Weight decay	0.0005
Gamma(ϒ)	0.1
Batch size	100

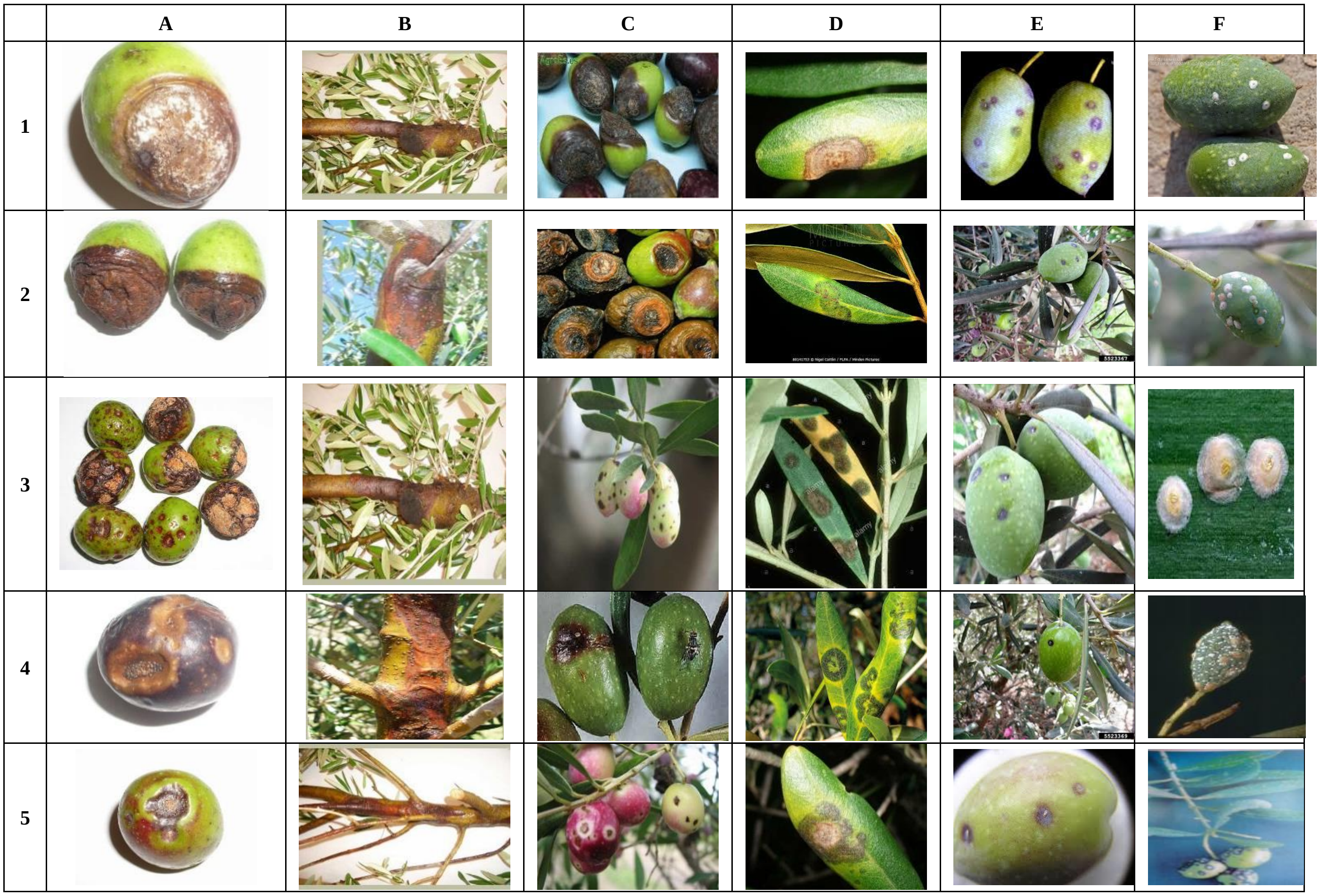


Fig. 5. Sample of Olive Leaf Diseases used in the Experiment. (A) Anthracnose, (B) Canker, (C) Lepra Fruit Rot, (D) Peacock Spot, (E) Parlatoria Oleae, (F) Aspidiotus Nerii.

VI. Results and Discussion

Using the proposed DL model network architecture, plant leaves are trained and classified to detect both crop species and disease identity in two main experiments. As mentioned above, one works at the original PlantVillage dataset while the other works at the enhanced dataset using the images acquired from JOUF lab. All experiments are configured to run for a total of 6 epochs each, and they consistently converge after a few steps down in the learning rate. As this paper plans to identify 14 olive diseases, we prefer to document only the results of the final experiment which takes olive Pkant into account. In order to validate the results of the proposed model, the renewed dataset is divided into 80% training and 20% testing sets. "Table II" shows the True Positive (TP), False Positive (TP), True Negative (TN), and False Negative (FN) respectively for olive diseases.

Detailed metrics (precision, recall, F1 Measure, sensitivity, and specificity) for each olive disease are depicted in "Table III". The minimum values obtained are highlighted in a red underlined font. ‘Peac’ has the lowest precision value; ‘Frost’ has both recall and F1 vlaue, while ‘Gloe’ has the lowest sensitivity. All specificity measure for all diseases is 100%.

TABLE. II. TP, FP, TN, and FN for Olive Diseases

Syndrome	**TP**	**FP**	**TN**	**FN**
Anth	72	0	2215	0
Aspid	144	2	2138	4
Cankers	143	5	2138	0
Frost	1	0	2283	1
Gloe	419	0	1867	0
Hail	7	1	2279	5
Lepra	129	2	2156	0
Macr	59	2	2226	0
Marciume	36	0	2246	0
Parl	6	1	2280	1
Peac	37	4	2245	3
Phyt	104	0	2183	1
Pseud	71	1	2211	3
Tuber	1	0	2286	0

TABLE. III. OLIVE DISEASES THAT ARE CONSIDERED BY THIS STUDY

Syndrome	Precision	Recall	F_1 Measure	Sensitivity	Specificity
Anth	100	100	100	100	100
Aspid	100	98.63	99.31	100	100
Cankers	97.95	99.31	98.62	98	100
Frost	100	50	66.67	99	100
Gloe	100	100	100	25	100
Hail	100	100	100	100	100
Lepra	100	100	100	100	100
Macr	100	100	100	100	100
Marciume	100	100	100	100	100
Parl	100	100	100	88	100
Peac	90.24	97.37	93.67	100	100
Phyt	100	100	100	97	100
Pseud	100	100	100	100	100
Tuber	100	98.63	99.31	100	100

"Fig. 6" alongside "Table IV" summarizes the remarked results in contrast to a number of the state-of-art methods. The proposed method achieves an overall accuracy about 99.11% which is the highest mark. Besides, it has 99.49%, 99.11%, and 99.29% in terms of precision, recall, and F1 measure respectively. These metrics are the highest as opposed to other methods. It has been noticed that our proposed model outperforms all other methods in terms of overall accuracy, precision, recall, and F1 measure. This singularity in measurements is a result of efficient artificial augmentation, balanced number of class images, optimal parameters assignment, smart network configuration, and usefulness of transfer learning. The problems arising in most traditional attempts [25-28] to detect plant diseases follow from hand-engineered features, image enhancement techniques, and labor-intensive methodologies. These traditional attempts are lean to either a small number images in a class or a limited variety of classes of crops. Unlike [17], the proposed model AlexNet model is trained over an enhanced version of PlantVillage dataset. Thus, it is more general for multiple disease identification for apple, tomato, and olive leaf diseases. Moreover, it relies on augmentation and achieved an overall accuracy of 99.11% for disease identification.

To conclude, it has been observed that the performance of DL models in image classification has made a remarkable progress in the past few years [29-31]. Previous traditional approaches such as SIFT [32], HoG [33], SURF [34], etc., and the likes were based on hand-engineered features extraction methods. These approaches heavily depend on predefined features. They also lack the transfer learning. In other words, they fail once the problem at hand is renewed or major changes are introduced to the dataset.

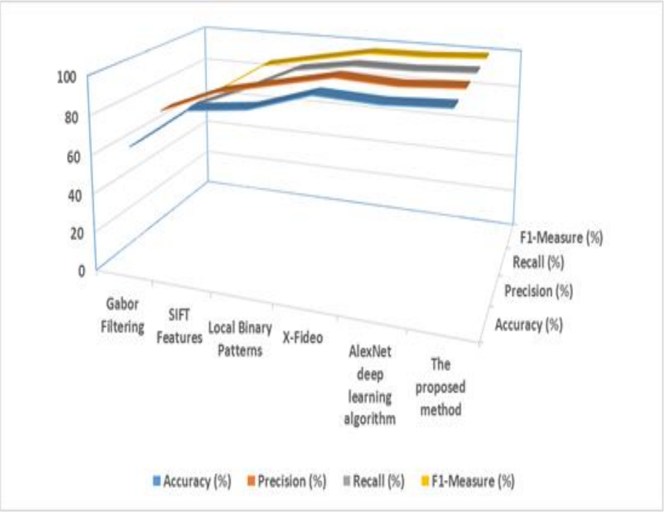


Fig. 6. The Proposed Method Compared to State-of-Art Methods.

TABLE. IV. PROPOSED METHOD COMPARED TO STATE-OF-ART METHODS

The proposed method	Accuracy (%)	Precision (%)	Recall (%)	F_1-Measure (%)
Background Suppressing Gabor Energy Filtering [26] with RBF-SVM [27].	63.11 ± 11.91	72.44 ± 14.30	65.28 ± 21.74	65.52 ± 15.15
SIFT Features[28] and RBF-SVM [27].	84.91 ± 17.44	85.19 ± 20.07	77.87 ± 43.61	84.65 ± 18.13
Uniform Local Binary Patterns [35] and RBF-SVM [27].	88.55 ± 16.71	92.12 ± 17.68	92.24 ± 6.16	90.95 ± 11.97
X-Fideo (LeNet deep learning algorithm) [17].	98.60 ± 1.47	98.82 ± 2.63	97.18 ± 2.71	96.89 ± 3.45
AlexNet deep learning algorithm [15].	97.38 ± 1.89	97.42 ± 1.33	97.37 ± 1.45	97.36 ± 2.45
The proposed method	99.11 ± 0.75	99.49 ± 0.83	99.11 ± 1.29	99.29 ± 1.63

VII. Conclusion and Future Work

Early detection of plant diseases has been taken as a positive move to maintain and enhance crop quality and reduce production loss to the minimum. As a result, DL approaches gain wide acceptance worldwide due to their accuracy and efficiency in plant disease detection field. The remarked results compared to a number of state-of-the-art methods are promising. Our proposed method achieves an overall accuracy of 99.11% which is the highest mark. Besides, it has 99.49%, 99.11%, and 99.29% in terms of precision, recall, and F_1 measure respectively. It outperforms all other methods when it comes to overall accuracy, precision, recall, and F_1 measure. More interestingly, although the model training consumes ample time, the classification during testing runs quickly in a few seconds even on a CPU. Therefore, the model could be easily implemented on a smartphone. In the future, a smartphone-assisted crop disease diagnosis will be targeted, and the proposed model will be available at Mobil apps.

Acknowledgment

We are indebted to the Deanship of Scientific Research at Jouf University for funding this work through General Research Project under grant number (39/766). We also extend our gratitude to Olive Research Center as well as Camel and Range Research Center for their support during our data collection of olive diseases and pests.

References

[1] Kaur, S., S. Pandey, and S. Goel, Plants disease identification and classification through leaf images: A survey. Archives of Computational Methods in Engineering, 2019. 26(2): p. 507-530.

[2] ABDULLAKASIM, W., et al., An images analysis technique for recognition of brown leaf spot disease in cassava. Tarım Makinaları Bilimi Dergisi, 2011. 7(2): p. 165-169.

[3] Bhange, M. and H. Hingoliwala, Smart farming: Pomegranate disease detection using image processing. Procedia Computer Science, 2015. 58: p. 280-288.

[4] Mohan, K.J., M. Balasubramanian, and S. Palanivel, Detection and recognition of diseases from paddy plant leaf images. International Journal of Computer Applications, 2016. 144(12).

[5] Phadikar, S., J. Sil, and A.K. Das, Rice diseases classification using feature selection and rule generation techniques. Computers and electronics in agriculture, 2013. 90: p. 76-85.

[6] Amara, J., B. Bouaziz, and A. Algergawy. A Deep Learning-based Approach for Banana Leaf Diseases Classification. in BTW (Workshops). 2017.

[7] Fuentes, A., et al., A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors, 2017. 17(9): p. 2022.

[8] Islam, M., et al. Detection of potato diseases using image segmentation and multiclass support vector machine. in 2017 IEEE 30th canadian conference on electrical and computer engineering (CCECE). 2017. IEEE.

[9] Atole, R.R. and D. Park, A multiclass deep convolutional neural network classifier for detection of common rice plant anomalies. International Journal of Advanced Computer Science and Applications, 2018. 9(1): p. 67-70.

[10] Ramcharan, A., et al., A mobile-based deep learning model for cassava disease diagnosis. Frontiers in Plant Science, 2019. 10: p. 272.

[11] Singh, U.P., et al., Multilayer Convolution Neural Network for the Classification of Mango Leaves Infected by Anthracnose Disease. IEEE Access, 2019. 7: p. 43721-43729.

[12] Tian, Y., et al., Detection of Apple Lesions in Orchards Based on Deep Learning Methods of CycleGAN and YOLOV3-Dense. Journal of Sensors, 2019. 2019.

[13] Jiang, P., et al., Real-Time Detection of Apple Leaf Diseases Using Deep Learning Approach Based on Improved Convolutional Neural Networks. IEEE Access, 2019. 7: p. 59069-59080.

[14] Al-Tarawneh, M.S., An empirical investigation of olive leave spot disease using auto-cropping segmentation and fuzzy C-means classification. World Applied Sciences Journal, 2013. 23(9): p. 1207-1211.

[15] Prasanna Mohanty, S., D. Hughes, and M. Salathe, Using Deep Learning for Image-Based Plant Disease Detection. arXiv preprint arXiv:1604.03169, 2016.

[16] Sladojevic, S., et al., Deep neural networks based recognition of plant diseases by leaf image classification. Computational intelligence and neuroscience, 2016. 2016.

[17] Cruz, A.C., et al., X-FIDO: An effective application for detecting olive quick decline syndrome with deep learning and data fusion. Frontiers in plant science, 2017. 8: p. 1741.

[18] Kour, V.P. and S. Arora, Fruit Disease Detection Using Rule-Based Classification, in Smart Innovations in Communication and Computational Sciences. 2019, Springer. p. 295-312.

[19] Goodfellow, I., Y. Bengio, and A. Courville, Deep learning. 2016: MIT press.

[20] Kingma, D.P. and J. Ba, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[21] Taylor, G., et al. Training neural networks without gradients: A scalable admm approach. in International conference on machine learning. 2016.

[22] Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems. 2012.

[23] LeCun, Y., et al., Backpropagation applied to handwritten zip code recognition. Neural computation, 1989. 1(4): p. 541-551.

[24] Han, X., et al., Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. Remote Sensing, 2017. 9(8): p. 848.

[25] Srivastava, N., et al., Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014. 15(1): p. 1929-1958.

[26] Cruz, A.C., B. Bhanu, and N.S. Thakoor, Background suppressing Gabor energy filtering. Pattern Recognition Letters, 2015. 52: p. 40-47.

[27] Chang, C.-C., " LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, 2: 27: 1--27: 27, 2011. http://www. csie. ntu. edu. tw/~ cjlin/libsvm, 2011. 2.

[28] Liu, C., J. Yuen, and A. Torralba, Sift flow: Dense correspondence across scenes and its applications. IEEE transactions on pattern analysis and machine intelligence, 2011. 33(5): p. 978-994.

[29] Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[30] Zeiler, M.D. and R. Fergus, Visualizing and understanding convolutional networks (2013). arXiv preprint arXiv:1311.2901, 2013.

[31] Szegedy, C., et al. Inception-v4, inception-resnet and the impact of residual connections on learning. in Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[32] Lowe, D.G., Distinctive image features from scale-invariant keypoints. International journal of computer vision, 2004. 60(2): p. 91-110.

[33] Dalal, N. and B. Triggs. Histograms of oriented gradients for human detection. in international Conference on computer vision & Pattern Recognition (CVPR'05). 2005. IEEE Computer Society.

[34] Bay, H., et al., Speeded-up robust features (SURF). Computer vision and image understanding, 2008. 110(3): p. 346-359.

[35] Almaev, T.R. and M.F. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. in 2013 Humaine Association Conference on.

Biotechnical System for Recording Phonocardiography

Marwan Ahmed Ahmed Hamid[1]
Dept. of Biomedical and Food Engineering
NED University of Engineering and Technology
Karachi, Pakistan

Maria Abdullah[2]
Dept. of Computer System Engineering, Dawood University
of Engineering and Technology, Karachi, Pakistan

Najeed Ahmed Khan[3]
Dept. of Computer Sc. and Info. Technology
NED University of Engineering andTechnology
Karachi, Pakistan

Yasmin Mohammed Ahmed AL-Zoom[4]
Dept. of science and Technology
IBB University, Ibb, Yemen

***Abstract*—The Phonocardiography is a graphical method of recording of the tones and noise generated by the heart with the help of the phonocardiogram machine. Cardiovascular disease (CVD) and heart failure (HF) are considered life-threatening and mostly cause death. The phonocardiograph signal (PCG) considers an indicator of abnormalities in the cardiovascular system. It provides the ability to carry out qualitative information and quantitative analysis of different tones and heart murmurs. PCG plays a major role in treatment, diagnosis and decision making of the clinical examination and biomedical research fields. The use of simple stethoscope for diagnosis the heart problem requires an experienced physician or doctors. Many people with CVD and HF are dying every day because of the lack of facilities that analysis the heart defects. Most of the low come countries suffer from a severe shortage in the Electrocardiogram (ECG) and PCG devices and trained physicians and doctors. The Poor healthcare system in these countries needs to be improved especially the problem of heart disease diagnostics. The PCG is a technique for recording and monitoring the cardiac acoustics by using a transducer and microphone. This paper attempts to design a cheap and simple biotechnical system for recording and monitoring PCG signal. The hardware was designed and implemented using the stethoscope, electret microphone, amplifiers, DC source, and jack for transmitting the PCG signal to the computer. The software was codded for monitoring and processing the PCG signal.**

Keywords—Phonocardiograph (PCG); cardiovascular diseases (CVD); heart failure (HF); electrocardiograph (ECG)

I. Introduction

The phonocardiography method is a non-invasive investigation method that is recommended to evaluate the HF and considered as an essential tool for diagnosing CVD [1]. It provides crucial information about several heart problems.

These days the PCG signal is less used because of the wide and effective use of the ECG signal and echocardiography, but still many cardiac defects are best diagnosed by the PCG [2]. The anatomy of the human heart is described as four chambers (two auricles and two ventricles) for collection and pumping out of the blood, respectively [3]. The heart beating (sound) is described as a complex interaction among pressure components, contraction of the heart chambers (muscles) and blood vessels flow [4]. The mechanical process of contraction of the heart muscles and the flow of the blood result in vibrations and acoustics that can be recorded over the chest [5].

The PCG is one of the most effective ways to evaluate heat function. Wang Haibin et al. [6] developed a system for in-home use of heart abnormality monitoring. Their system can auscultate respiratory sounds, heart sounds and lung sounds. Therefore, these sounds contain a lot of valuable information for the diagnosis of heart problems.

Recently, with rapid growth of computer hardware and digital signal processing methods, PCG signal could be recorded and analyzed [7]. A. Atbi el al [8] presented an algorithm for detection of the heart sound. Their method based on remove the heart noise by using the low pass filters and also detected the S1 and S2 peaks of the heart sound.

The recorded sound of the heart (PCG) can be defined as the main (S1 and S2) and sub main (S3 and S4) components [9]. Fig. 1 shows the all classes of PCG components whereas Fig. 2 shows location of S-S interval of PCG and R-R interval of ECG.

The diastolic segment (S1and S2) is important in cardiovascular diagnosis. The Diastolic part of the PCG signal studied Metin Akay et al. [10] they identified the S1 as the loudest class of PCG signal and it happens during ventricular contraction. The S1 class contains a series of low frequency vibrations ranging between 25Hz and higher amplitudes with time terms of 100ms to 200ms. The S2 class appears at the end of ventricular systole for about 0.12s with a frequency of 50Hz, which is higher in frequency, but shorter in time terms than S2 [11]. The third class (S3) can be heard at the beginning of diastole, while a class S4 in the PCG signal [12].

The improvement of the biomedical engineering technology has led to renew some of the useful medical detection methods. Noor Kamal et al. [13] introduce new approach to detect specific measurements that relate to biomedical signal such as PCG and ECG. In their result, the first and second heart sound (S1, S2) can be determined from another signal like ECG.

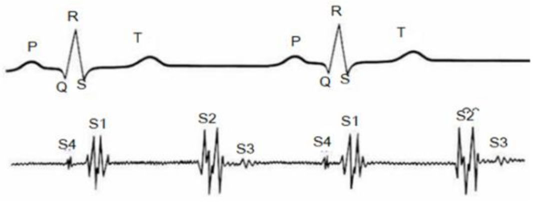


Fig. 1. All Classes (S1, S2, S3, and S4) of the PCG Signal.

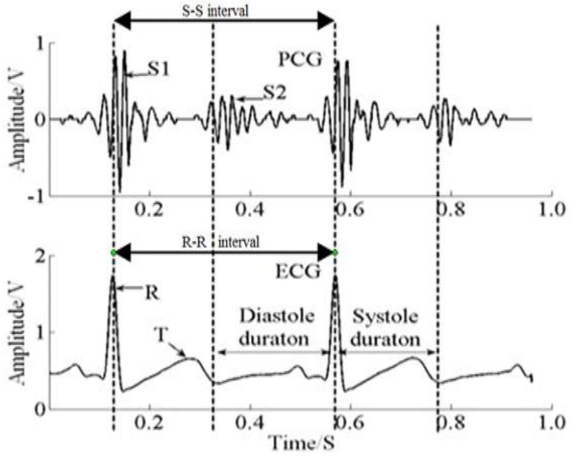


Fig. 2. Main Components of Normal PCG Signal with Location of S-S Interval of PCG and R-R Interval of ECG.

The rate of cardiovascular diseases with each passing day is alarming [14] especially in poor countries. The importance of the phonocardiography can be identified from the simplicity of this technique and also from the amount of information that can have from PCG signal. The doctors and physicians mostly use the heart sound to evaluate the different heart functions and detect the abnormalities of the heart. The shortage of the medical diagnostic equipment is the main reason for the poor diagnostics and then treatment of the heart diseases. The following methodology has been employed to overcome the problem of the heart medical diagnostics by developing an efficient and uncostly biotechnical system for recording and storing the PCG signal.

II. Proposed Methodology

Auscultation of the heart is a method of listening to the sound of the heart and can be recorded from specific points on the chest. In medical diagnostics, the PCG signal is collected from the S1, S2, S3, and S4 components over the chest as demonstrated in Fig. 3.

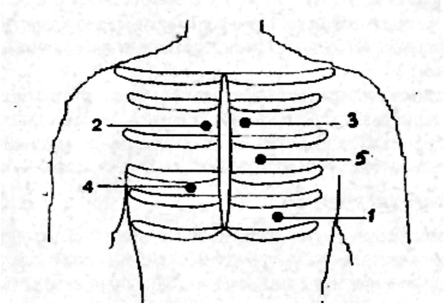


Fig. 3. Heart Auscultation Points.

In this paper we proposed a biotechnical system that may be a modified stethoscope for recording the PCG signal. In proposed biotechnical system the earpieces of the stethoscope are replaced by an electret (sensor (CZN-15E) microphones. The output from the sensors is amplified by amplifiers (NE5534P). The jack connector fed the amplified signal and then fed into PC system. The recoded PCG signal undergoes several steps of transmitting the sound into an electrical signal, then the electrical signal is amplified and converted to a digital signal [15]. The signal process is preformed to denoise and improve the quality of the PCG signal. The software has a friendly graphical user interface (GUI) for monitoring the PCG signal [16]. The data and metadata can be stored in digital form. The proposed biotechnical system contains the following parts.

A. Selection of Components and Circuit Design

The selection of appropriate elements for any circuit is important to get an accurate and efficient result. The selection of the stethoscope (Fig. 4) is not much important, as the selection of an electret microphone Fig. 5. The electret microphone is a type of capacitor microphone that does not need a power supply for polarizing voltage [17]. The proposed electret microphone in this method is a sensor type of CZN-15E. Table I presents the characteristics of the required electret microphone (CZN-15E).

Fig. 4. Stethoscope.

Fig. 5. Electret Microphone.

TABLE. I. CHARACTERISTIC OF THE SELECTED ELECTRET MICROPHONE (CZN-15E)

Feature	Description
Sensitivity	-58±2dB (0dB=1V/pa,1KHz)
Impedance	Low impedance
Directivity	Omnidirectional
Frequency	20-16,000Hz
Voltage range	1.5V-10V
Standard operation voltage	4.5V
Current consumption	Max.0.5mA
Sensitivity reduction	Within -3dB at 3V
S/N ratio	More than 60dB

The heart sound hits the diaphragm by vibrated air particles, as a result, the diaphragm changes the distance between the plates. The electret substance moves on the backplate, producing a voltage across the capacitor [18]. The produced voltage is weak and needs to be fed to the amplifier. The required amplifier should be low noise with high-speed audio. The implemented operational amplifier in this method is NE5534P [19]. Table II contains the main features of the required amplifier (NE5543P).

The electrical circuit diagram was designed via any appropriate circuit design software (Multisim [20]). Fig. 6 shows the biotechnical circuit of the proposed system.

TABLE. II. CHARACTERISTIC OF THE SELECTED AMPLIFIER (NE5534P)

Feature	Description
Equivalent Input Noise Voltage	3.5 nV/√Hz Typ
Unity-Gain Bandwidth	10 MHz Typ
Common-Mode Rejection Ratio	100 dB Typ
High DC Voltage Gain	100 V/mV Typ
Peak-to-Peak Output Voltage Swing	32V Typ with VCC± = ±18 V and RL = 600 Ω
High Slew Rate	13 V/μs Typ
Wide Supply-Voltage Range	±3 V to ±20 V
Low Harmonic Distortion	
Offset Nulling Capability and External Compensation Capability	

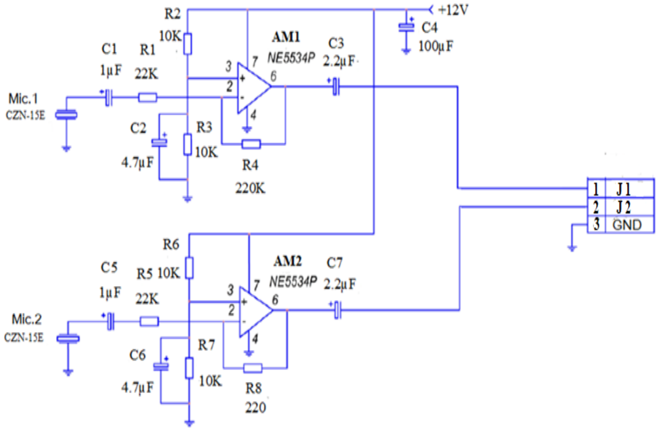


Fig. 6. The Circuit Diagrams for the Proposed Biotechnical System.

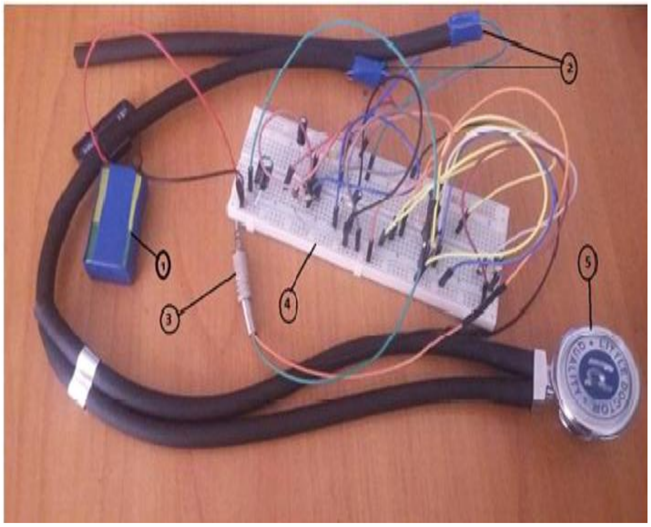


Fig. 7. Proposed Biotechnical System for Recording PCG. Numeric Indicates 1- DC Power Supply; 2- Microphone System; 3-Connector; 4- Transmitter unit; 5- Stethoscope.

The design of the biotechnical system (hardware) is shown in Fig. 7 consists of:

1) Two Electret microphones system type CZN-15E) and two amplifiers type NE5534P

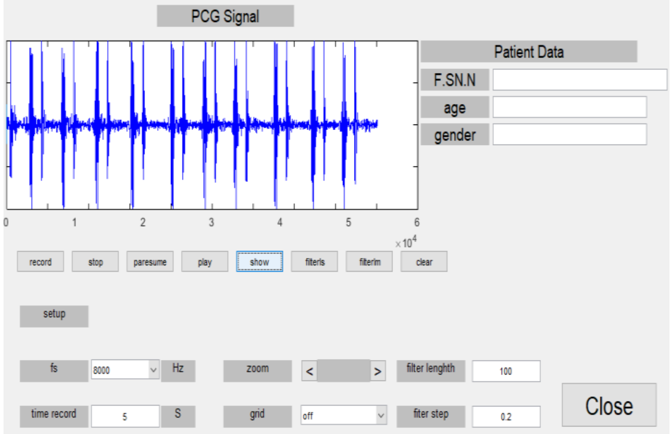
2) Transducer block and connector (3.5 jack) for transmitting the signal to the computer

3) DC Power supply (12V)

B. The Description of the Software Interface

The interface of the proposed system is shown in Fig. 8. The setup section (Fig. 9) contains parameters as follows:

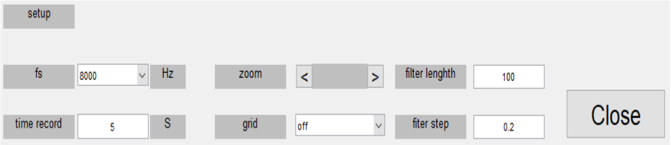
- Sampling frequency (Fs) is the number of samples per second in a PCG signal [21].
- Time recoding (S) for controlling the length of the PCG signal.
- The resizing the length of the windowing filter is done by filter length button.
- In the PCG interface special button for zooming the PCG signal.



Patient Data
F.SN.N
age
gender

fs 8000 Hz
time record 5 S
grid off
filter lenghth 100
fiter step 0.2

Fig. 8. Graphical user Interface of the Biotechnical.



setup

fs 8000 Hz

time record 5 S

zoom < >

grid off

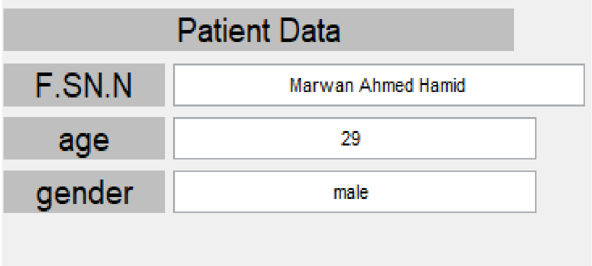
filter lenghth 100

fiter step 0.2

Fig. 9. Setup Section.

Fig. 10 illustrates the patient information section. In this section, the patient's metadata such as name, age, and gender can be registered and saved.

In Fig. 11 to manipulate and control the PCG signal undesired signals (noises) are expected to be recorded parallel with the PCG signal due to the physical movement of the patients or the around environment. The button of filtration contains low pass filter to remove different noises.



Patient Data

F.SN.N Marwan Ahmed Hamid

age 29

gender male

Fig. 10. Section of the Metadata.

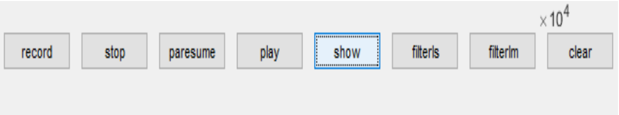


Fig. 11. Recording and Controlling Buttons.

III. Result and Discussion

By the methodology mentioned above, the proposed biotechnical system was able to record the PCG signal (Fig. 12). The S1 and S2 peaks present clearly in the recorded PCG signal. Experimentally, it has been noticed that S1 peak values are normally larger than S2 values. Fig. 13 illustrates the time gab between S1 and S2 peaks.

The recorded PCG signal in Fig. 13 contains different type of noises [22]. In this system we used digital signal processing to denoise and enhance the quality of signal. Fig. 14 shows the PCG signal after filtering and smoothing.

The output of proposed system is plotting high fidelity recording of the heart sound. The reliability of the obtained results is verified by doctors.

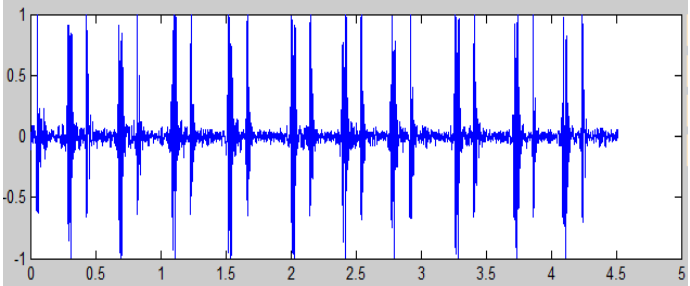


Fig. 12. Recorded PCG Signal.

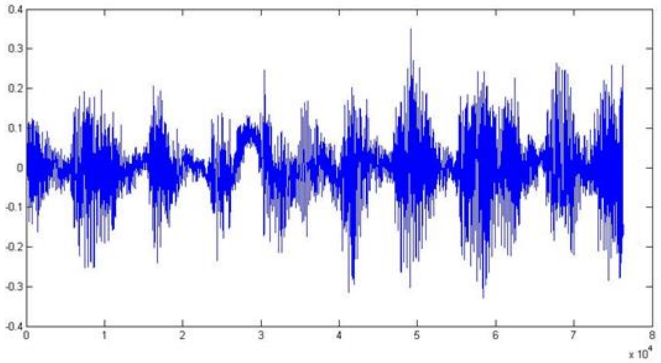


Fig. 13. S1 and S2 Peaks of the PCG Signal before Filtering and Smoothing.

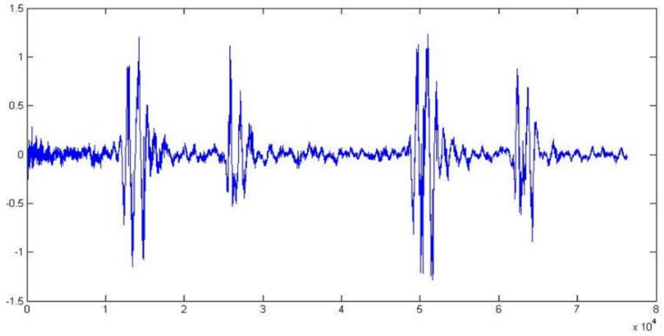


Fig. 14. PCG Signal after Denoising.

IV. Conclusion

The stethoscope requires an experienced physician or doctors to analysis the heart sound. The PCG provides a useful method for studying and a diagnosis of the heart sound. By following the steps mentioned above, heart sound can be experimentally recorded and stored. The graphical user interface GUI for PCG biotechnical system (software-based system) allows easy interaction with an electrical circuit (hardware). Practically the S1 and S2 peak can be graphically detected. The confirmation of obtained results can be done by the doctors. The results (data) can be used and stored easily on a computer. The proposed method contributes to future heart sound recording research. The improvement in accuracy of the system will be better with comparative and stronger components (Hardware). The software (code) can be used in cellphones and research on signal processing methods enhances the quality of obtained PCG signal.

Confliect of Interest

The authors confirm that this article content has no conflict of interest.

References

[1] Studies and E. Engineering, “Detection of Heart Diseases by Mathematical Artificial Intelligence Algorithm Using Phonocardiogram Signals,” vol. 3, no. 1, pp. 145–150, 2013.

[2] P. Pcg and S. Recognition, “A thesis submitted in conformity with the requirements Graduate Department of Electrical and Computer Engineering Abstract,” 2009.

[3] J. Sundberg-cohon, “The Heart: Anatomy, Physiology and Exercise Physiology,”. January, 2009.

[4] L. S. Lilly, “The Cardiac Cycle : Mechanisms of Heart Sounds and Murmurs,” pp. 29–45.

[5] L. H. Cherif, “Segmentation of heart sounds and heart murmurs,” January, 2016.

[6] H. Wang, Y. Fang, Z. Gao, Y. Wang, and H. You, “A Heart Sound Acquisition and Analysis System Based on PSoC4,” vol. 7, no. 2, pp. 22–31, 2017.

[7] D. Theodor and K. Gretzinger, “Analysis of Heart Sounds and Murmurs by Digital Signal Manipulation by l,” 1996.

[8] A. Atbi and S. M. Debbal, “ Segmentation of Pathological Signals by Using Shannon Energy Envelogram,” vol. 2, no. 1, pp. 1–14, 2013.

[9] T. Chowdhury, “Phonocardiography and analysis using standard deviation profile A major qualifying thesis report :,” no. January, 2017.

[10] Y. M. Akay, “Detection of Coronary Occlusions Using Autoregressive Modeling of Diastolic Heart Sounds,” no. May, 1990.

[11] N. Dia, J. Fontecave-jallon, P. Gumery, and B. Rivet, “Denoising Phonocardiogram signals with Non-negative Matrix Factorization informed by synchronous Electrocardiogram,” pp. 51–55, 2018.

[12] R. R. Sarbandi, J. D. Doyle, M. Navidbakhsh, K. Hassani, and H. Torabiyan, “A color spectrographic phonocardiography (CSP) applied to the detection and characterization of heart murmurs : preliminary results,” Biomed. Eng. Online, vol. 10, no. 1, p. 42, 2011.

[13] N. K. Al-qazzaz, I. F. Abdulazez, and S. A. Ridha, “Simulation Recording of an ECG , PCG , and PPG for Feature Extractions,”. November, 2015.

[14] G. Son and S. Kwon, “applied sciences Classification of Heart Sound Signal Using Multiple Features,” 2018.

[15] A. Mirzal, “Principles of Signal Conversion : A Brief Tutorial,”. January, 2018.

[16] E. C. Microphone, “Guide for Electret Condenser Microphones.”

[17] R. Article, “Short review of devices for detection of human breath sounds and heart tones,” vol. 6, no. 3, 2014.

[18] J. Hillenbrand, S. Haberzettl, and G. M. Sessler, “Electret microphones with stiff diaphragms,”. January, 2016.

[19] A. Equipment, C. Circuits, T. C. Amplifiers, and M. Equipment, “SE5534A Single Low Noise Operational Amplifier Pb − Free Packages are Available,” pp. 1–10, 2012.

[20] U. Guide, “Electronics Workbench TM Multisim TM 9 Simulation and Capture User Guide,” . February, p. 794, 2006.

[21] A. Jeukendrup, “Heart rate monitoring: applications and limitations,” no. June, 2017.

[22] C. S. Sim, J. H. Sung, S. H. Cheon, J. M. Lee, J. W. Lee, and J. Lee, “The Effects of Different Noise Types on Heart Rate Variability in Men,” vol. 56, no. 1, pp. 235–243, 2015.

A Guideline for Decision-making on Business Intelligence and Customer Relationship Management among Clinics

Nur Izzati Yusof[1], Norziha Megat Mohd. Zainuddin[2],
Noor Hafizah Hassan[3], Nilam Nur Amir Sjarif[4],
Suraya Yaacob[5]
Advanced Informatics Department, Razak Faculty of
Technology and Informatics, Universiti Teknologi
Malaysia, 54100 Kuala Lumpur, Malaysia

Wan Azlan Wan Hassan[6]
Computing Department, Faculty of Communication
Visual Art and Computing
Universiti Selangor, Bestari Jaya, 45600
Selangor
Malaysia

***Abstract*—Business intelligence offers the capability to gain insights and perform better in decision-making by using a particular set of technologies and tools. A company's success to a certain extent depends on customers. The complementary of business intelligence and customer relationship management will improve the efficiency of organizations, hence increase productivity and revenue. Most research works on implementation of business intelligence and customer relationship management in organizations commonly concentrate on architecture, framework, and maturity model. The process on how to implement business intelligence and customer relationship management in an organization, especially in smaller domains has not yet been clarified which make some organizations unclear on how to implement business intelligence and customer relationship management. Thus, this study investigates the process involved in the implementation of business intelligence and customer relationship management among clinics. An infographic guideline was developed based on the six process of data mining which is known as Cross Industry Standard Process for Data Mining. Four elements of business intelligence decision-making process which were gather, store, access, and analyze were also included in the process of developing the guideline. Findings from an expert's review show that the increase of Content Validity Index was 0.7, from 0.3 during the first iteration to 1.0 in the second iteration. Therefore, this result is acceptable. The guideline appears to be a useful instrument for practitioners to implement business intelligence and customer relationship management in their clinics, however the process involved in developing the guideline could be improvised from time to time.**

Keywords—Business intelligence; customer relationship management; decision-making; guideline

I. Introduction

Business intelligence (BI) is seen as the technology and method for managing data in order to enhance decision-making [1]. BI technology provides businesses to gather, store, access and analyze huge volume of information in order to make better customer, supplier, employee, logistics, and infrastructure choices [2]. Decision maker in organizations will get benefits in making decision when implementing BI. Previous researchers have considered BI as a process and product [3]. The process consists of techniques used by organizations to create helpful data, or intelligence, which can assist organizations to survive and flourish. The product is data that will enable organizations to predict with a degree of certainty the behavior of their rivals, providers, clients, technologies, acquisitions, markets, goods and services, and the overall companies' climate. The advancement of BI system is significant in many industries, and this is no exception in healthcare.

In recent years, many studies on BI and CRM implementation in healthcare have been conducted and they only concentrated on larger domains that are generally exposed to the advancement of technology [4]. There are three commonly used models in implementing BI and CRM solution which are architecture, framework and maturity model. Thus, literatures on BI and CRM implementation in smaller organizations are required. The knowledge deficit on BI and CRM among practitioners in small organizations must be considered thoroughly. To address this issue, this study proposes a guideline for decision-making in implementing BI and CRM system among clinics. This guideline is useful to assist clinic practitioners in making better decision by fully utilizing the system.

Several attributes were considered in developing the guideline. Two employed instruments were questionnaire and guideline. Four elements of BI decision making process, six steps from cross-industry standard process for data mining (CRISP-DM) and two principles of infographics [5]–[7] were adapted in developing the guideline. Face validity and content validity were used to assess the questionnaire and the guideline. The result of evaluation was measured by using Context Validity Index (CVI). The purpose of this study is to address the issue on decision-making of practitioners in implementing BI and CRM system among clinics. More specifically, this study has two objectives:

1) To develop a guideline for decision-making in a clinic that can facilitate the clinician to fully utilize the implementation of BI and CRM.

2) To evaluate the developed guideline for decision-making in a clinic that can facilitate the clinician to fully utilize the implementation of BI and CRM.

This article is organized into several sections. The immediate section that follows provides background information on business intelligence, customer relationship management, and infographics which clarify the domain-specific needs in the healthcare industry. Section III presents the methodology employed in the development of the guideline. Section IV presents the discussion on the findings. Finally, Section V presents the conclusion of this study by summarizing the contribution, limitation, and recommendation for future research.

II. Literature Review

The definitions made by researchers on business intelligence vary. According to Gartner Group (2012), BI can be defined as a comprehensive form of applications and technologies for gathering, storing, analyzing, sharing, and providing access to data to support companies to make better business decisions [8]. As mentioned by Gartner, BI helps decision makers to make better decision-making by transforming data into beneficial knowledge. Apart from that, BI also is defined as an environment where business users can easily and intuitively obtain consistent, reliable, perceptible and manipulated data to help managers search for organizational knowledge in the past, present and future [9]. From these definitions, it can be said that BI provides knowledge to decision makers in order to help them in making better decision regarding their business.

Customer relationship management (CRM) is an overall process of developing and maintaining lucrative customer relationships by providing superior customer value and satisfaction [10]. The word customer relationship management includes all the ideas that businesses use in relation to their customers, including the collection, storage and evaluation of client data, while considering the privacy and safety of the data [11]. Several researchers have suggested that CRM is applied to preserve the loyalty of customers as well as to increase profits. This is also applicable to healthcare organizations in which they can foster patient-clinic connection and provide more tangible advantages [12]–[14]. The researchers believe that the implementation of CRM system in healthcare is very important. This is because an efficient CRM system integrates individual health records, patient records and hospital information in order to provide a solution to manage health-related problems, advantages and expenses [15].

Studies on integration of BI and CRM systems have shown that this integration has created a path to customer loyalty [16]–[18]. Researchers have emphasized that the complementary use of CRM systems and business intelligence offers a holistic advantage that involves the enhancement of client profiling, simplifying client detection value, and evaluating the company's achievement in satisfying customers. BI can enhance incentives such as quicker transformation of potential into real customers, reduced amount of outgoing customers and increased sales among current customers. Therefore, in the modern business, CRM cannot be considered separately from business intelligence. These two develop a distinctive model that allows companies to predict customers' behaviors and make decisions based on these predictions and eventually, build long-term and profitable customer relationships [19]–[21].

The discussions on delivering information differ among studies conducted. Several studies have found that data visualization helps audiences to digest information better [22]. The use of infographics generates attention and interest as it utilizes words, numbers, icons, colors, and graphics that can be used in a more focused, organized, intuitive and engaging to tell a story behind the information and data [23]. Infographic is derived from the word information and graphic. It is used to disclose specific information to specific users. Infographic can be a tool for better understanding of knowledge, as it provides wider spectrum of information in visual form to audiences. It supports audiences to discover deeply and provides valuable facts that are hidden in complexity. Therefore, infographic can be described as a graphic design that combines data visualization, illustrations, text and image, together into a format that tells a complete story [24].

III. Methodology

This study applies a quantitative research methodology. A set of questionnaires was distributed to end users for the evaluation of the proposed guideline. The development of the guideline was based on the four elements of business intelligence decision-making concept. The elements were gather, store, access and analyze [2]. In addition, the process in developing the guideline was stimulated by the six processes of CRISP-DM [6]. Both the questionnaire and guideline were evaluated by experts to determine their usability.

A. Respondents

As suggested by Nielsen, three to five evaluators are enough to avoid more problems [25]. In this case study, five respondents, three experts and one end user were selected to participate in the validation and evaluation of the proposed guideline. The respondents and the experts were students and senior lecturers in a public university, while the end user was an owner of two dental clinics. The experts were chosen based on their expertise which was applicable to this case study. The list of experts and their expertise can be seen in Table I. The dental clinic was chosen because this clinic did not use any integration system with its other branch. This would cause difficulties in monitoring patients in clinical management. To add to that, the owner of this clinic was unaware of BI and CRM system. Therefore, the proposed guideline is hoped to assist decision makers in gathering more awareness and insights into the significance of having BI and CRM in order to improve the clinics' services.

TABLE. I. The Specialization and Departments of Experts

No.	Experts	Specialization	Department
1.	Expert 1	Business Intelligence	Advanced Informatics
2.	Expert 2	Visualization and Interactive Design	Advanced Informatics
3.	Expert 3	Information Systems, Human-Computer Interaction (HCI)	Advanced Informatics

B. Research Instruments

This study employed two instruments which were guideline and questionnaire. Fig. 1 shows the development of the instrument employed.

Fig. 1 shows that there are two types of instrument used in this study; they were a guideline and a questionnaire. The guideline had undergone one phase of validation which was on content validity. Meanwhile, the questionnaire had undergone two phases of validation which were face validity and content validity. Three different participants were involved in the validation and evaluation process. Five respondents were selected to participate in the face validity phase in which they were asked to validate whether the questionnaire could be understood or otherwise. Three experts were selected to participate in the content validity assessment of both the questionnaire and guideline. The changes in the content on questionnaire and the guideline were made according to their evaluation and validation. The last participant who was involved in the validation and evaluation measurement was an end user. The end user validated the usability and the likeliness of the guideline based on the given questionnaire. The Likert scale from strongly disagree to strongly agree was used to measure the acceptance of the guideline.

To develop an infographic design of the guideline, the data presentation and visualization tool was used. In this study, Visme, a free online tool that can create various types of infographic was used to design the guideline. The development of the guideline involved three components that must be considered as illustrated in Fig. 2.

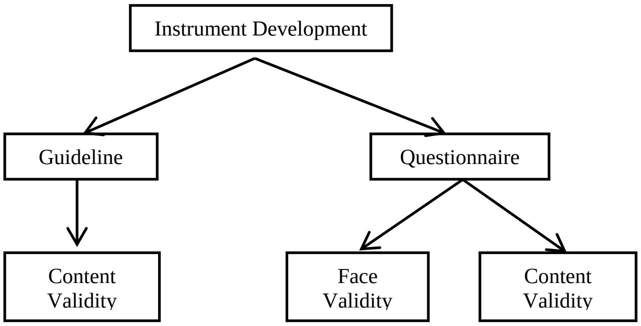


Fig. 1. Instrument for Decision-Making Guideline.

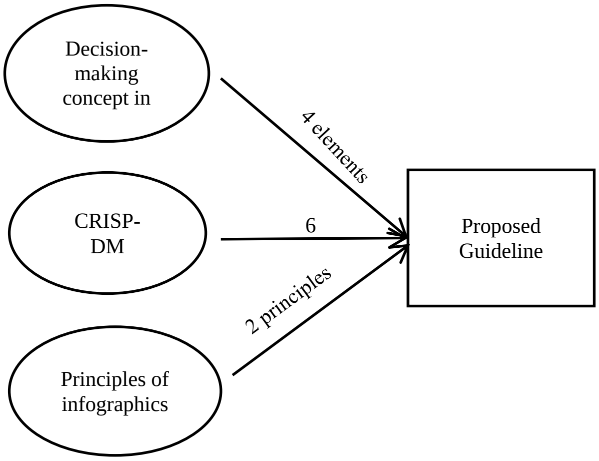


Fig. 2. Guideline Development.

According to Fig. 2, the proposed guideline was developed based on four elements of decision-making concept in BI, six processes of CRISP-DM and the principles of infographics [2], [6], [7]. Four elements of decision-making process in BI included (i) gather, (ii) store, (iii) access, and (iv) analyze. The six processes of CRISP-DM were business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Meanwhile, the two principles of infographics [7] are:

1) Process infographic: This type of infographic shows the step by step process by visualizing the number and flow of the process in a connected line so that audience can quickly understand a process.

2) Informational infographic: This type of infographic shows information to audience. This infographic provides extra information to audiences by adding a descriptive header so that audience can understand the infographic better.

The proposed guideline consisted of six processes as can be seen in Table II. The processes were grouped into four elements of decision-making process in BI. The elements were used in the questionnaire to differentiate the processes of decision-making that were relevant to those elements so that the respondents would able to comprehend BI and CRM implementation better. Table III shows the items in the questionnaire.

The proposed guideline contains of six processes adopted from CRISP-DM [6]. They are:

1) Understand your business: A clear objective and the obstacles found in the clinic must be clearly identified.

2) Understand your data: Collect data of patients and drugs inventories as well as other clinic management matters.

3) Prepare your data: The data of the patients, drugs inventories, clinic tools and other related management matters must be recorded accurately. Irrelevant, old and redundant data has to be removed and cleaned.

4) Adopt online database system: Store and update the data in an online database so that the data is integrated and can be easily accessed anywhere and anytime.

5) Assess the data: Access the data to identify any problem occurred and overcome the problem as soon as possible.

6) Monitor, maintain and iterate usage: Regular assessment on current business performance to improve decision-making.

TABLE. II. SIX PROCESSES OF THE GUIDELINE

Number of Process	**Description**	**Element of BI**
1.	Understand your business problem	Gather
2.	Understand your data	
3.	Prepare your data	Store
4.	Adopt online database system	
5.	Assess the data	Access
6.	Monitor, maintain and iterate usage	Analyze

TABLE. III. THE ITEMS OF QUESTIONNAIRE

Item	Items	Reference
Gather		
GI	The terms used in these steps is easy to understand and can be followed.	[26]
G2	I must identify the objective of my clinic clearly.	[27]
G3	I must identify any limitation occurred before I proceed to my future project plan.	[27]
G4	The medical data (including patients record, drugs inventories, etc.) must be well-recorded.	Self-developed
G5	Patients satisfaction towards the clinic services is my priority.	Self-developed
Store		
S1	The terms used in these steps is easy to understand and can be followed.	[26]
S2	Medical data (including patients record, drugs inventories, etc.) should be recorded accordingly to its category.	[27]
S3	All redundant or same data have to be removed as well as old and irrelevant data.	[27]
S4	If I store the medical data in an online database system, it will help me to access and analyze the medical record anywhere and anytime.	[27]
Access		
Ac1	The terms used in this step is easy to understand and can be followed.	[26]
Ac2	The record of patients' data has to be evaluated to encounter any issues involved (i.e.: illness, allergic, bill, total patients visit).	Self-developed
Ac3	The record of drugs inventories has to be assessed to aware any low and waste stocks.	Self-developed
Ac4	Cash flow of the clinic is important to measure the financial availability.	[27]
Ac5	It is important to assess patients' feedbacks and satisfaction.	Self-developed
Analyze		
An1	The terms used in this step is easy to understand and can be followed.	[26]
An2	It is important to measure clinic performance regularly.	[27]
An3	Current business report is important in decision-making to ensure better management performance in future.	[27]
An4	Patients' experience and satisfaction is one of the most keys to success.	[28]
An5	An adequate service will bond a good relationship with patients to preserve their loyalty.	Self-developed
Overall Evaluation		
E1	The title and the content of the guideline is understandable.	[26]
E2	The guideline presents a complete, clear, well-formed message and is logically structured.	[26]
E3	The graphic of the guideline includes relevant text, images and consistent design elements.	Self-developed
E4	The guideline enables me to improve my knowledge on decision-making of customer relationship management.	Self-developed

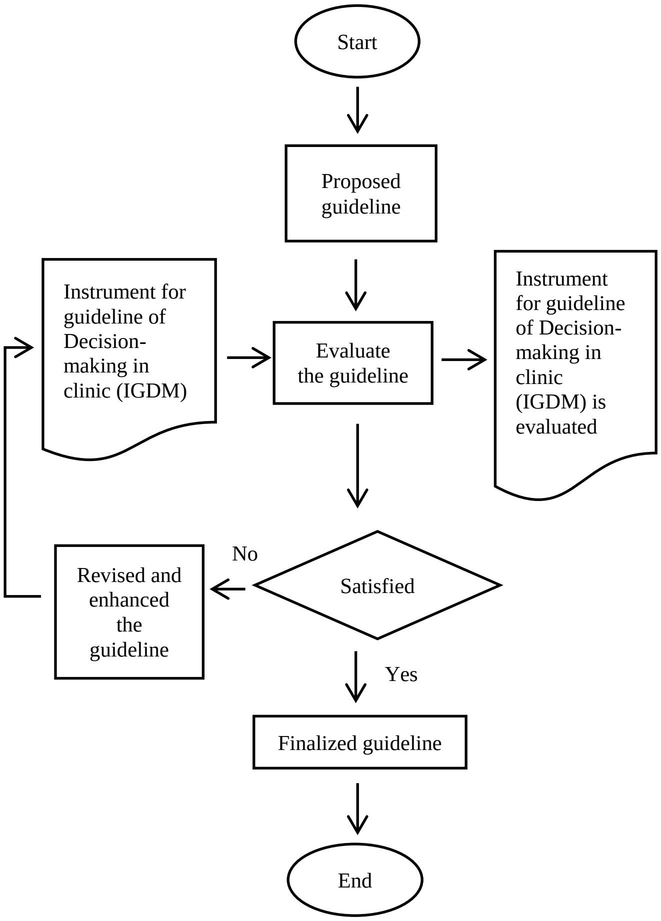

Start
Proposed guideline
Instrument for guideline of Decision-making in clinic (IGDM)
Evaluate the guideline
Instrument for guideline of Decision-making in clinic (IGDM) is evaluated
No
Revised and enhanced the guideline
Satisfied
Yes
Finalized guideline
End

Fig. 3. Flowchart of Guideline Evaluation.

C. Research Procedures

Before the guideline can be employed, evaluation processes must be conducted first. Fig. 3 shows the flowchart of the evaluation of the guideline.

The process started with proposing a guideline. The guideline was evaluated by employing face validity and content validity tests. Face validity was used to identify whether the term and sentences used could be easily understood. The content validity was used to validate the content of the guideline in terms of graphics and information that was included in the guideline. Once the guideline was approved, it can be used. Otherwise, the guideline had to be revised and improvised. The revised guideline will then repeat the evaluation process until it would approve and appropriate to be used.

IV. FINDINGS

The experts were asked to validate the content of the questionnaire and the infographic guideline. In doing this, the instrument was sent to experts to be validated and they were modified based on the comments made by the experts. Three experts were selected for the content validity. The first expert had visualization expertise, the second expert had business intelligence expertise, and the third expert had questionnaire expertise. Table IV provides the summary on the comments made by the experts.

TABLE. IV. EXPERTS' COMMENTS ON CONTENT VALIDITY

No.	Experts	Comments
1.	Expert 1	a. The question is too general. Try to precise and deep further the element of infographic in line with the element of questionnaire. b. The flow of the information is according to the process need to be taken – clarity of the process through information flow. c. Include cue in the infographic by using metaphor/analogy/gestalt. d. Emphasized right keyword. Increase font size. Reduce irrelevant picture.
2.	Expert 2	a. The question is more on infographic. Add more element of BI in the questionnaires. b. Restructure the flow of the guideline. Try to follow the process of CRISP-DM. c. Some content of the guideline is not clear. d. Relate the questions to BI decision-making/decision-making in clinic.
3.	Expert 3	a. Some wording needs to be revised so that the meaning can reflect the situation better. b. Add extra explanation on BI terms to give understanding for audience who does not have information technology (IT) background.

TABLE. V. CONTENT VALIDITY INDEX (CVI) – BEFORE REVISED

Questions/ Items	Expert 1	Expert 2	Expert 3	No. of Agreement	I/CVI
Q1	2	1	2	0	0
Q2	2	4	2	1	0.333333
Q3	4	2	4	2	0.666667
Q4	2	4	4	2	0.666667
Q5	2	4	4	2	0.666667
Q6	4	1	2	1	0.333333
Q7	2	1	2	0	0
Q8	2	1	2	0	0
Q9	2	1	2	0	0
Q10	2	4	4	2	0.666667
Q11	2	4	2	1	0.333333
Q12	2	2	2	0	0
Q13	2	2	4	1	0.333333
Q14	2	2	2	0	0
Q15	2	1	2	0	0
Q16	2	2	2	0	0
Q17	2	2	2	0	0
Q18	2	2	4	1	0.333333
Q19	4	2	4	2	0.666667
Q20	4	2	2	1	0.333333
Q21	4	2	2	1	0.333333
Q22	4	2	4	2	0.666667
Q23	4	2	4	2	0.666667
				S/CVI	0.304348

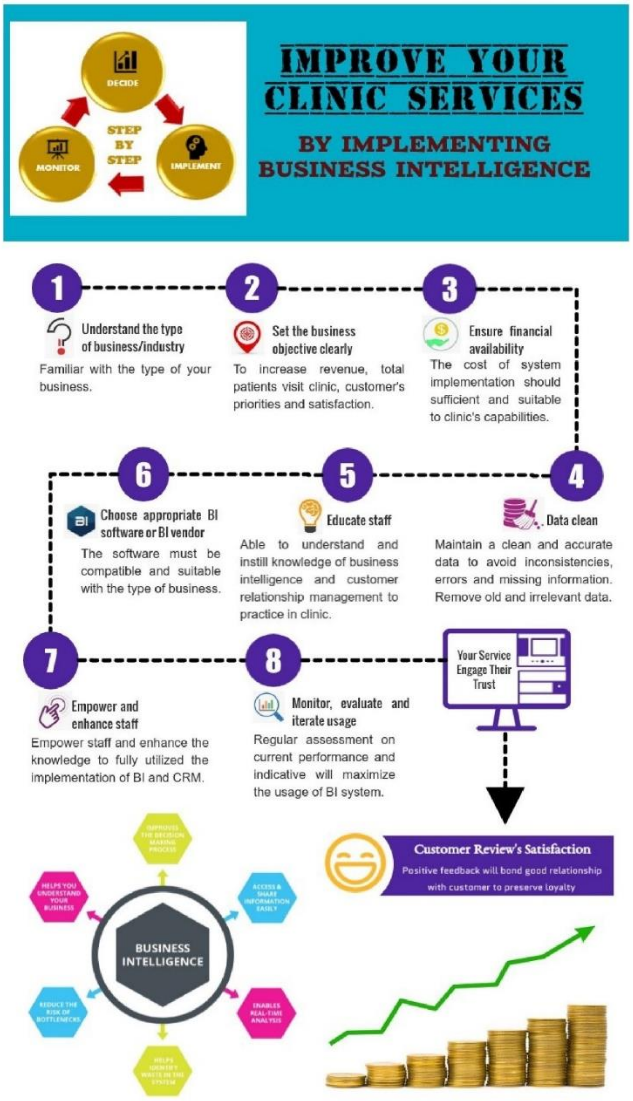
In order to evaluate the expert result on content validity of the questionnaire, the Content Validity Index (CVI) was used in this study. Content validity is the degree to which an instrument has an appropriate sample of items for the construct being measured and is an important procedure in scale development. CVI is the most widely used index in quantitative evaluation [29]. A CVI value can be computed for each question on a scale which refers to I-CVI, and the overall scale which refers to S-CVI. In this study, the experts were asked to rate the relevance of each item, on a 4-point scale. The scale used in this study was 1=Not Relevant, 2=Somewhat Relevant, 3=Relevant, 4=Highly Relevant. Researchers recommend that a scale with excellent content validity should be composed of I-CVIs of 0.78 or higher and S-CVI/UA and S-CVI/Ave of 0.8 and 0.9 or higher, respectively. The results of the CVI on pre and post revision are depicted in Table V and Table VI.

Table V and Table VI show the results of Content Validity Index. Before the questionnaire was revised, the computed result of S/CVI was 0.304348. This result expressed that the questionnaire needed a lot of improvement. After the questionnaire was revised, the computed result of S/CVI was 1. This proved that the revised questionnaire was accepted to be used.

TABLE. VI. CONTENT VALIDITY INDEX (CVI) – AFTER REVISED

Questions/ Items	Expert 1	Expert 2	Expert 3	No. of Agreement	I/CVI
Q1	4	4	4	3	1
Q2	3	4	4	3	1
Q3	4	3	3	3	1
Q4	3	4	4	3	1
Q5	4	4	4	3	1
Q6	4	4	3	3	1
Q7	4	3	4	3	1
Q8	4	4	4	3	1
Q9	3	4	4	3	1
Q10	4	4	3	3	1
Q11	4	4	4	3	1
Q12	3	3	4	3	1
Q13	4	4	4	3	1
Q14	4	4	3	3	1
Q15	3	4	4	3	1
Q16	4	4	4	3	1
Q17	4	4	3	3	1
Q18	4	3	4	3	1
Q19	4	4	4	3	1
Q20	3	4	4	3	1
Q21	4	3	4	3	1
Q22	4	4	3	3	1
Q23	3	4	4	3	1
				S/CVI	1

Based on the experts' review, the infographic guideline needed to be revised so that the audience could gain better understanding. The title, picture, the steps, and the font were the elements that needed to be revised. The results of the guideline pre and post revision are depicted in Fig. 4 and Fig. 5.


DECIDE
IMPLEMENT
MONITOR
STEP BY STEP
IMPROVE YOUR CLINIC SERVICES
BY IMPLEMENTING BUSINESS INTELLIGENCE
1
2
3
4
5
6
7
8
Your Service Engage Their Trust
BI
BUSINESS INTELLIGENCE
IMPROVES THE DECISION MAKING PROCESS
HELPS YOU UNDERSTAND YOUR BUSINESS
ACCESS & SHARE INFORMATION EASILY
REDUCE THE RISK OF BOTTLENECKS
ENABLES REAL-TIME ANALYSIS
HELPS IDENTIFY WASTE IN THE SYSTEM

Understand the type of business/industry
Familiar with the type of your business.

Set the business objective clearly
To increase revenue, total patients visit clinic, customer's priorities and satisfaction.

Ensure financial availability
The cost of system implementation should sufficient and suitable to clinic's capabilities.

Data clean
Maintain a clean and accurate data to avoid inconsistencies, errors and missing information. Remove old and irrelevant data.

Educate staff
Able to understand and instill knowledge of business intelligence and customer relationship management to practice in clinic.

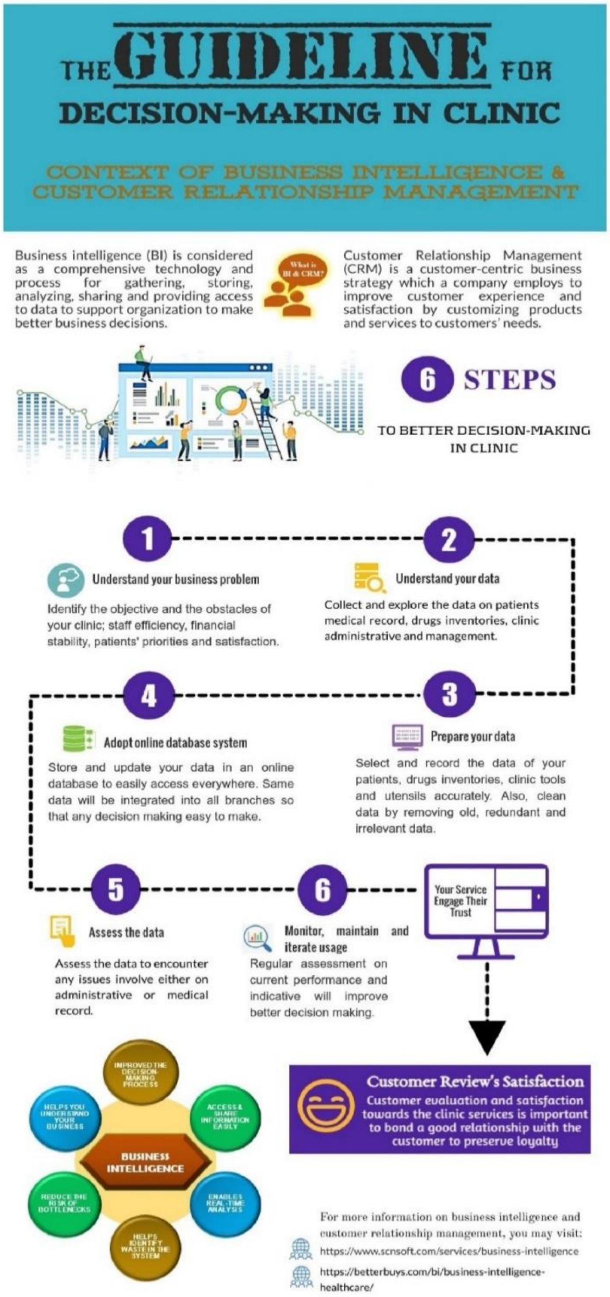
Choose appropriate BI software or BI vendor
The software must be compatible and suitable with the type of business.

Empower and enhance staff
Empower staff and enhance the knowledge to fully utilized the implementation of BI and CRM.

Monitor, evaluate and iterate usage
Regular assessment on current performance and indicative will maximize the usage of BI system.

Customer Review's Satisfaction
Positive feedback will bond good relationship with customer to preserve loyalty

Fig. 4. Infographic Guideline I–before Revised.


THE GUIDELINE FOR
DECISION-MAKING IN CLINIC
CONTEXT OF BUSINESS INTELLIGENCE & CUSTOMER RELATIONSHIP MANAGEMENT
What is BI & CRM?
6 STEPS
TO BETTER DECISION-MAKING IN CLINIC
1
2
3
4
5
6
Your Service Engage Their Trust
BUSINESS INTELLIGENCE
IMPROVED THE DECISION-MAKING PROCESS
HELPS YOU UNDERSTAND YOUR BUSINESS
ACCESS & SHARE INFORMATION EASILY
REDUCE THE RISK OF BOTTLENECKS
ENABLES REAL-TIME ANALYSIS
HELPS IDENTIFY WASTE IN THE SYSTEM

Business intelligence (BI) is considered as a comprehensive technology and process for gathering, storing, analyzing, sharing and providing access to data to support organization to make better business decisions.

Customer Relationship Management (CRM) is a customer-centric business strategy which a company employs to improve customer experience and satisfaction by customizing products and services to customers' needs.

Understand your business problem
Identify the objective and the obstacles of your clinic; staff efficiency, financial stability, patients' priorities and satisfaction.

Understand your data
Collect and explore the data on patients medical record, drugs inventories, clinic administrative and management.

Prepare your data
Select and record the data of your patients, drugs inventories, clinic tools and utensils accurately. Also, clean data by removing old, redundant and irrelevant data.

Adopt online database system
Store and update your data in an online database to easily access everywhere. Same data will be integrated into all branches so that any decision making easy to make.

Assess the data
Assess the data to encounter any issues involve either on administrative or medical record.

Monitor, maintain and iterate usage
Regular assessment on current performance and indicative will improve better decision making.

Customer Review's Satisfaction
Customer evaluation and satisfaction towards the clinic services is important to bond a good relationship with the customer to preserve loyalty

For more information on business intelligence and customer relationship management, you may visit:
https://www.scnsoft.com/services/business-intelligence
https://betterbuys.com/bi/business-intelligence-healthcare/

Fig. 5. Infographic Guideline II–after Revised.

Based on Fig. 5, it can be observed that the title, the steps, and the picture were the elements that were revised. From the guideline (before revised) the title from "Improve Your Clinic Service by Implementing Business Intelligence" was been changed to "The Guideline for Decision-Making in Clinic, Context of Business Intelligence and Customer Relationship Management" (after revised). The irrelevant picture was also removed. For example, the picture of the coins did not fit the research. Infographic Guideline II had included more information on the definition of BI and CRM, and also the given link for audience to know more about BI and CRM. The expert asked to put the link for easy access among the audience to know more about BI and CRM. Apart from that, the most important criteria of the guideline were the steps involved in decision-making. From the experts' review, instead of providing steps to implement BI; which was too general, the expert recommended to change the steps on decision-making according to the CRISP-DM process. Therefore, this study has followed the steps involved in CRISP-DM and has been slightly revised to fit the context of the research. Based on Infographic Guideline I, there were 8 steps involved to implement the BI and CRM. Meanwhile, after the revision, only 6 steps were used in Infographic Guideline II in line with the steps included in CRISP-DM. The summary of the changes before and after the guideline was revised and it can be seen in Table VII.

TABLE. VII. SUMMARY OF INFOGRAPHIC GUIDELINE'S CHANGES ITEM

No.	Item	Before Changes	After Changes
1.	Title	Improve Your Clinic Service by Implementing Business Intelligence.	The Guideline for Decision-Making In Clinic, Context of Business Intelligence and Customer Relationship Management.
2.	Steps	8 steps.	6 steps.
3.	Info	Lack of BI and CRM information.	Include the information of BI and CRM.
4.	Picture	Irrelevant picture that does not related to the element of BI and CRM.	The irrelevant picture has been removed.

V. DISCUSSION AND CONCLUSIONS

This study identified three attributes of developing guideline for decision-making to implement BI and CRM among clinics. The attributes contained four elements of decision-making concept in BI, six process adapted from CRISP-DM process and two principles of infographics. This guideline is useful in assisting practitioners to make decisions by implementing BI and CRM. Face validity and content validity were used to evaluate the guideline. CVI shows the result of 1 after experts' reviews. The infographic guideline was also improvised. The decision-making to implement BI and CRM is important nowadays as previous studies have also emphasized the importance of implementing BI and CRM which would improve the performance of business [13], [14], [30]–[32].

The implementation of BI and CRM is expanding each day. Many large companies have implemented the BI and CRM with the proper standard. Yet, small companies are still way behind in implementing BI and CRM and lack expertise to fully utilize it. In this project, a clinic has been selected as a case study. Since the population is increasing, the need of improving health care management is important to help people with different types of illness to be treated well. Since technology is advancing, the efficiency of operation and management of clinics need to be aligned with the latest technology. However, without a proper guideline to help decision makers to decide carefully for the sake of the performance of their clinics, the use of high technology will give a little impact to them. Therefore, a guideline for decision-making on CRM among clinics is proposed to assist decision makers in making better decisions. It is an advantage for them to know better the use of BI and CRM.

Time consumption contributes one of the main limitations of this study. The challenge was to find a clinic that wanted to participate since many clinics were too occupied with their daily operations and handling patients, not many were ready to participate due to their tight schedules. Apart from that, another limitation was finding the clinics that fit into the perimeter of this study. Much time was spent on making phone calls and paying visits that these premises to find the ones that were suitable for this study.

The proposed guideline is recommended to be used in various type of industries, instead of just focusing only on health care institutions. The purpose is to improve the reliability of the results and allowing the research to represent various industries. It is also hoped that this study will further add to the pool of knowledge on BI and CRM and opens more doors for similar research works to be conducted in the future.

REFERENCES

[1] P. Wanda and S. Stian, "The Secret of my Success : An exploratory study of Business Intelligence management in the Norwegian Industry," Procedia - Procedia Comput. Sci., vol. 64, no. 1877, pp. 240–247, 2015.

[2] C. Lennerholt, J. van Laere, and E. Söderström, "Implementation Challenges of Self Service Business Intelligence: A Literature Review," Proc. 51st Hawaii Int. Conf. Syst. Sci., vol. 9, pp. 5055–5063, 2018.

[3] N. Caseiro and A. Coelho, "The influence of Business Intelligence capacity, network learning and innovativeness on startups performance," no. 2017, 2018.

[4] L. Gastaldi, A. Pietrosi, S. Lessanibahri, M. Paparella, A. Scaccianoce, G. Provenzale, M. Corso, and B. Gridelli, "Technological Forecasting & Social Change Measuring the maturity of business intelligence in healthcare : Supporting the development of a roadmap toward precision medicine within ISMETT hospital," Technol. Forecast. Soc. Chang., vol. 128, no. August 2017, pp. 84–103, 2018.

[5] A. Ahmed, "Exploration of business intelligence using Oralce B.I (OBIEE)," 2017. [Online]. Available: https://www.slideshare.net/AneelAhmed/exploration-of-business-intelligence-using-oralce-bi-obiee.

[6] H. Kemper, H. Baars, and H. Lasi, "Business Intelligence and Performance Management," Bus. Intell. Perform. Manag., pp. 13–27, 2013.

[7] S. Mcguire, "9 Types of Infographics and When to Use Them [+ Infographic Templates]," 2019. [Online]. Available: https://venngage.com/blog/9-types-of-infographic-template/. [Accessed: 15-Jun-2019].

[8] O. Ali, P. Crvenkovski, and H. Johnson, "Using a Business Intelligence Data Analytics Solution in Healthcare A case study : Improving Hip Fracture Care Processes in a Regional Rehabilitation System," 2016.

[9] S. Williams and N. Williams, The Profit Impact of Business Intelligence. 2010.

[10] A. Sen and A. P. Sinha, "IT alignment strategies for customer relationship management," Decis. Support Syst., vol. 51, no. 3, pp. 609–619, 2011.

[11] C. Mati and L. Ilie, “Customer relationship management in the insurance industry,” vol. 15, no. 14, pp. 1138–1145, 2014.

[12] P. L. H. C.L. Hsu, C. Chiu, “‘A Constraint-Based Optimization Mechanism for Patient Satisfaction,’ Knowledge-Based Intelligent Information and Engineering Systems.” 2014.

[13] Y. M. Baashar, “An Integrative Perspective for CRMS Implementation in Healthcare in Malaysia,” no. 100, 2014.

[14] M. N. Almunawar and M. Anshari, “Improving Customer Service in Healthcare,” no. August 2018, 2011.

[15] M. Boris, “Application Of Customer Relationship Management Strategy (Crm) In Different Business Areas Boris Milovic,” vol. 9, pp. 341–354, 2012.

[16] A. Habul, “Business Intelligence and Customer Relationship Management,” pp. 169–174, 2010.

[17] S. R. Shinde and Sunjita, “Integration between Customer Relationship Management and Business Intelligence,” 2018.

[18] A. Khan, N. Ehsan, E. Mirza, and S. Zahoor, “Integration between Customer Relationship Management (CRM) and Data Warehousing,” vol. 1, pp. 239–249, 2012.

[19] Y. Gupta and N. Sharma, “When BI Meets CRM: An Emerging Concept in Retail Industry,” nternational J. Bus. Anal. Intell. New Delhi, vol. 1, no. 1, pp. 41–48, 2013.

[20] A. Habul, A. Pilav-Velić, and K. Emir, “Customer Relationship Management and Business Intelligence,” Intech, vol. i, no. tourism, p. 13, 2016.

[21] R. Gujrati, “CRM for retailers: Business intelligence in retail CRM,” vol. 2, no. 1, pp. 24–29, 2016.

[22] M. Lonsdale and D. Lonsdale, Information Visualization Guidelines. 2019.

[23] K. T. Lyra, S. Isotani, R. C. D. Reis, L. B. Marques, L. Z. Pedro, P. A. Jaques, and I. I. Bitencourt, “Infographics or Graphics+Text: Which material is best for robust learning?,” Proc. - IEEE 16th Int. Conf. Adv. Learn. Technol. ICALT 2016, pp. 366–370, 2016.

[24] N. S. Arum, “Infographic: Not Just a Beautiful Visualisation,” Univ. Birmingham, 2017.

[25] C. W. Turner, J. R. Lewis, and J. Nielsen, “Determining Usability Test Sample Size,” vol. 3, no. 2, pp. 3084–3088, 2006.

[26] J. C. Dunlap and P. R. Lowenthal, “Getting graphic about infographics : design lessons learned from popular infographics,” vol. 6529, no. September, 2016.

[27] G. R. Gangadharan and S. N. Swami, “Business Intelligence Systems: Design and Implementation Strategies,” 26th Int. Conf. Inf. Technol. Interfaces ITI, pp. 139–144, 2004.

[28] H. Han and S. S. Hyun, “Customer retention in the medical tourism industry: Impact of quality, satisfaction, trust, and price reasonableness,” Tour. Manag., vol. 46, pp. 20–29, 2015.

[29] Z. N. D. X. X. Bao and Y. X. Ban, “Content validity index in scale development,” 2012.

[30] Y. Li, J. Huang, and T. Song, “Information & Management Examining business value of customer relationship management systems : IT usage and two-stage model perspectives,” Inf. Manag., no. July, pp. 1–11, 2018.

[31] H. Moghimi, E. Healthcare, S. Vaughan, E. Healthcare, S. Mcconche, and E. Healthcare, “How Do Business Analytics and Business Intelligence Contribute to Improving Care Efficiency ?,” pp. 3408–3415, 2016.

[32] D. Coelho, J. Miranda, F. Portela, J. Machado, and M. Filipe, “Towards of a Business Intelligence Platform to Portuguese Misericórdias,” Procedia - Procedia Comput. Sci., vol. 100, pp. 762–767, 2016.

Cognitive Neural Network Classifier for Fault Management in Cloud Data Center

S.Indirani[1]
Research Scholar
Department of Computer Science
Mother Teresa Women's University, Kodaikanal-624102

Dr.C.Jothi Venkateswaran[2]
Former Associate Professor and Head
PG and Research Department of Computer Science
Presidency College (Autonomous), Chennai-600 005

***Abstract*—Pro-actively handling the fault in data center is a means to allocate the VM to Host before failures, so that SLA meets for the tasks running in the data center. Existing solution [1] on fault prediction in datacenter is based on a single parameter of temperature and the fault tolerance is implemented as a reactive solution in terms of VM replication. Different from these works, a proactive fault tolerance with fault prediction based on deep learning with multiple parameters is proposed in this work. In this work Cognitive Neural Network (CNN) is used to predict the failure of hosts and initiate migration or avoid allocation to the hosts which has high probability of failures. Hosts in the data center are scored on failure probability (FP-Score) based on parameters collected at various levels using CNN. VM placement and migration policies are fine-tuned using FP-Score to manage the failure proactively.**

Keywords*—Deep learning; Cognitive Neural Network (CNN); FP-Score; fault tolerance; VM allocation; VM migration

I. Introduction

Cloud data center has become a low cost solution for hosting of users computation and storage due to its unlimited resources on demand and pay as go model. Increasing number of users and enterprises are migrating their storage and computations to cloud data centers. Host failures happen in data center due to various reasons like memory exhaustion, hardware failures, software failures etc. Even in case of host failures ensuring the continuity of user's tasks with full or partial recovery in least latency is an important fault tolerance characteristic. Without it the quality of service of the data center is reduced and it will impact the business of data center provider.

Proactive fault tolerance is a way to reduce the down time and ensure higher QOS for the data center. It involves prediction of VM or host failures in advance and corrective actions to avoid the failure or in case of failure, reduce the impact of failure. Proactive fault tolerance has the overhead of increased resource consumption but considering reactive fault tolerance, it reduces the downtime and for the increased QOS provided, the overhead is reasonable. Deep Learning models are the latest trend in machine learning. They are used for various predictions in different domains of economic, health care, weather forecast and manufacturing etc. Deep learning models have the capability to learn the high quality features and semantic relations automatically from the structured and unstructured information compared to previous machine learning models where features are selected manually.

In this work host fault is modeled using CNN. The features collected across all of the OSI and OS layer at each sampling period is used as input for the prediction model. From the error logs collected over a sampling period from Cloud and OS layer a fuzzy failure probability score (FP-Score) is calculated. Training set is created labeling the OSI and OS layer features with FP-Score. CNN model is trained to predict the FP-Score from the Cloud and OS metrics. The VM Placement and VM migration policies in data center are modified in accordance with FP-Score. Dynamic replication strategy is also proposed for VM in certain cases to reduce the impact of failure.

II. Related Work

In [1] fault tolerant scheduling to enhance the reliability of tasks is proposed. The sub reliability requirements of tasks are calculated and from it the VM reliability is calculated. For critical task VM, reliability is ensured with replication.CPU temperature based failure prediction is proposed in [2]. The prediction function model for CPU temperature in the data center is modeled as

$$f(t|A, \omega, t_i, t_{i+1}) = \begin{cases} e^t, \quad 0 \le t \le t_i \\ e^{t_i}, \quad t_i \le t \le t_{i+1} \\ A\sin(\omega t - \omega t_{i+1}) + e^{t_i}, t_{i+1} \le t \le t_{i+2} \end{cases}$$

Based on the temperature model, the deteriorating physical machines are identified and the VMs are migrated from the rest of the physical machines selected using PSO based optimization. The time to recover from failure is very short in case of temperature based failure prediction. In [3] ranking based framework is proposed to locate significant components in cloud applications and rank them to provide fault tolerance for those components. The cloud application is modeled as weighted directed graph with each component as vertex and the component invocation as edges. The weight value of the edge is calculated based on invocation frequency.

$$W(e_{ij}) = \frac{frq_{ij}}{\sum_{j=1}^{n} frq_{ij}}$$

The significance value of the component (or node) is calculated as

$$V(C_i) = \frac{1-d}{n} + d \sum_{k \varepsilon N(c_i\ )} V(C_k)\, W(e_{ki})$$

Where n is the number of components, $N(c_i\ )$ is the set of components invoking C_i . The significance value above a threshold is determined as component which requires reliability and those components are replicated to ensure reliability. Fault prediction based on Byzantine fault detection is proposed in [4]. The fault model is constructed using service resource usage model and message transmission model of the cloud applications. Byzantine fault tolerance techniques are widely employed for building reliable systems but the parameters for modeling based on resource usage alone is not sufficient in data centers. The reliability of each host is modeled in terms of host ability to complete the task successfully in [5]. The task migration from low reliability node to high reliability is done for proactive fault tolerance. In [6] VM health status is monitored frequently and compared against QOS parameters to detect violations. Based on the violations, the risk level of VM is calculated. Load is shifted from high risk VM to low risk to provide increased reliability. A online predictive model for job failure using statistical learning techniques in proposed in [7]. Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Logistic Regression (LR) are applied for statistical learning. Time varying patterns of job failures and their system dependency is used to create a failure prediction model of tasks. The model is only used for aborting tasks which are predicted to fail and avoid the resource wastage. But the idea of predicting failures from statistical modeling of job attributes and system attributes can be used for migration decisions. The use of logs for error diagnosis is explored in [7]. Failure of system can be predicted based on the error logs. In [8] simple machine learning based on sparse coding is used to identify anomaly conditions. The model is trained for normal class data alone instead of both normal and abnormal class data. Normal class data captures run time behavior of a job that has not experienced a performance fault. The features deviation from normal class data is diagnosed as anomaly. The work in [9] proposed a anomaly prediction model for hosts based on parameters like CPU consumption, memory usage, input/output data rate, buffer queue length. Totally 20 features are collected in every 10 seconds to create a measurement stream. Triple state stream classifier was designed to classify each measurement sample to normal, alert or anomaly state. Prediction model triggers alarm whenever an alert state precedes anomaly state. An unsupervised non-intrusive domain independent framework for predicting latent faults in host is proposed in [10]. The main idea behind this method is to compare the machines performing the same task at same time. A machine deviating from normal behavior is tagged as anomalous. The performance counter values for machine m over period t is denoted as $x(m,t)$. To tag the machine m as suspicious, its value must be measured against other machines x values that run the same task at same time. To predict latent fault three tests sign test, turkey test and LOF test is proposed. Sign test is based on measuring $v(m)$ and termed as genuine when $v(m)$ is close to empirical mean of all machines.

$$v(m) = \frac{1}{T(M-1)} \sum_{t \in T} \sum_{m' \in M} \frac{x(m,t) - x(m',t)}{||x(m,t) - x(m',t)||}$$

Turkey test is based on calculation of Turkish depth function which gives high scores to points which are at center positions in the sample and low scores to points which are in the perimeter.LOF test is based on Local Outlier Factor outlier detection algorithm. The LOF function tries to find outliers by only looking at local neighborhoods. The assumption is that on different areas, the density of the sample might be different and this does not imply that points in less dense areas are outliers. The score of the LOF function is not calibrated. The greater the LOF score is, the more suspicious the point is. Based on this the suspicious hosts are identified. This approach is applicable only for speculative executive environment where resource consumption is not importance and only reliability is a concern. But in the proposed solution for fault tolerance resource consumption is also important. In [11] critical events identifying faults are determined and ranked based on their impact on fault diagnosis. The sequence of events in their order of occurrence that leads to fault is modeled in terms of a detection graph. Edge ranking algorithm is used the select the events critical to a particular fault and an event fault pattern is created. Critical events are selected based on three factors of affinity, weight and time decay of event. Affinity describes the involvement of event. Weight describes the discriminative power of the event and time decay how recent is the event to fault. The rank is calculated as

$$Rank(e) = \sum U_e \times W_e \times D_e$$

A sample event detection graph for three faults f1, f2, and f3 is given below in Fig. 1.

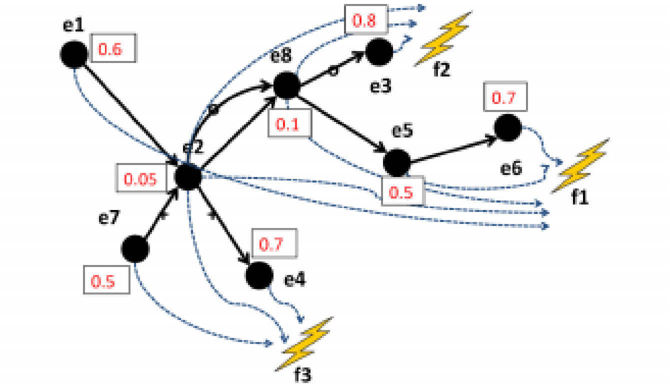


Fig. 1. Fault Graph [11].

The approach is adaptive to learn new fault patterns. But in typical datacenter, the number of events is very large and modeling the relationship is a cumbersome process. Redundant and irrelevant event removal can be done to reduce the modeling complexity. An automatic classification and identification of crisis in datacenter using efficient representation of datacenter state called finger print is proposed in [12].The finger print is learnt using statistical selection and summarization of the hundreds of performance metrics in datacenter. Performance metric in a particular epoch across all the application servers is summarized by computing the quantiles of the measured values (such as the median of CPU utilization on all servers). Based on the past observation, the values are characterized as hot, cold or normal. From it a summary vector containing one element per quantile per tracked metric, indicating whether the value of that quantile is cold, normal, or hot during that epoch is created. The summary vector is epoch finger print. Consecutive epoch finger prints are summarized as crisis fingerprint which characterizes the state of the datacenter. By matching to known crisis fingerprint, the state of datacenter can be predicted. The take way from this approach is that finger print can be done for host state and based on it task migration policy can be designed. An approach for fault localization in host by analyzing the dependencies among anomalous key performance indicators is proposed in [13]. A model is trained to capture the normal system behavior and this model is used to pinpoint the faulty resources that are likely to be responsible for possible failure. System execution under normal conditions in the form of relations among Key Performance Indicators (monitored KPIs), which are metrics collected on a regular basis from target resources of the system at different abstraction levels is used to train the model. A KPI base line model is created for normal executions. The deviation from base line model is measured frequently for tasks in hosts and task deviating greater than threshold is predicted to fail.

Unsupervised Behavior Learning (UBL) system for predicting failures in cloud is proposed in [14].UBL leverages an unsupervised learning method called the Self Organizing Map (SOM) which is able to capture complex system behavior with less expensive computational needs. System level behavior captured at OS and cloud layers are used to train SVM to classify normal behaviors. Deviation from normal system behaviors are predicted as anomalies. The continuous state of system from pre-failures to failure is modeled using SOM.UBL can raise an advanced alarm when the system leaves the normal state but has not yet entered the failure state. But for large datacenter, the application of UBL is more resource intensive. In [15] an integrated online anomaly prediction and virtualization-based prevention technique is proposed for virtualized cloud systems. Statistical learning is applied over system level metrics (cpu, memory, io statistics) for advance anomaly prediction and coarse grained anomaly cause inference in terms of identification of faulty VMs. An anomaly prediction model is created for each.

VM applies Tree Augmented Naïve Bayesian Network. It is an extension of the naive Bayesian model with consideration of dependencies among attributes into account. It can classify a system state into normal or abnormal and give a ranked list of metrics that are mostly related to the anomaly. The problem with this method is that momentary surge in VM load is also detected as fault without consideration for next state of VM.

III. Proposed Solution

Different from previous solutions, the proposed solution formulates a deep learning based failure prediction model. The model is able to predict the failure in advance such that there is enough time for migration and mitigate the failures. With cognitive neural learning model based on features extracted across all layers, the fault prediction capability is increased in the proposed solution. Each host is ranked in terms of FP-Score (values 1 to 10). Highest value denotes host is approaching failure sooner than lower value hosts. Features extracted across three categories care used for deep learning. The features collected at various layers are given in Table 1. The features are sampled at every window interval of time. Every t samples is aggregated in form of a 2-D matrix.

$$M = \begin{pmatrix} f11 & \cdots & f1N \\ \vdots & \ddots & \vdots \\ ft1 & \cdots & ftN \end{pmatrix}$$

TABLE. I. Features

Category	Name	Description
Cloud Layer	Average VM Memory utilization	The average memory utilization of VMs running in the host
	Average VM Disk Utilization	The average disk utilization of VMs running in the host
	Average VM Bandwidth Utilization	The average bandwidth utilization of VMs running in the host
	Normalized Cloudlet arrival rate	The task arrival rate at host
	CloudletSatisfaction Ratio	The rate at which incoming tasks are translated to cloudlet on a VM in the host
OS Layer	Host CPU Utilization	The utilization percentage of CPU in the host
	Host Memory Utilization	The utilization percentage of memory in the host
	Host Disk Utilization	The utilization percentage of Disk in the host
	Host Network Utilization	The utilization percentage of network bandwidth in the host
	Host temperature	The temperature of the host
	Host Resource Depletion Ratio Vector	The resource depletion ratio in terms of CPU and memory and disk.
Error Features	Cloudlet Error Histogram	The distribution count of cloudlet errors in log over various error levels.
	VM Error Histogram	The distribution count of VM errors in log over various error levels.
	OS Error Histogram	The distribution count of OS errors in log over various error levels.

Where N features collected over t sample intervals.

The matrix is labeled with a Failure Probability Score (FP).

A Loop controlled Cognitive Neural Network is trained to obtain the label for the input matrix M. Deep learning classifier is shown in Fig. 2. The architecture of the model is shown in Fig. 3. The cognitive network is a three layered feed forward radial bias function network. The transformation applied in each layer is given below in Table II.

The prediction output of the neural network classifier with K hidden layer neuron is given as

$$y_j^i = \propto_{j0} + \sum_{k=1}^{K} \propto_{jk} \Phi_k(X^i), j = 1,2 \dots n$$

Where the details of parameters are in Table 3.

The parameter for fine tuning the performance of the cognitive neural network is done by Loop control.

A CNN classifier is trained to predict the output label (FP-Score) for any input sample of Cloud, OS, and Error features expressed in form of 2-D matrix. With this classifier, the FP-Score of all hosts in data center is calculated periodically. The VM allocation and migration policies in the data center are modified based on the FP-Score. FP-VM allocation and FP-VM migration are the two policies proposed in this work integrating FP score of host with allocation and migration decisions.

FP-VM allocation does not create all VMs at once. Instead VMs are created in a progressive manner. Task are allocated to VM with best fit and in case a match is not found, task are kept pending in queue. The hosts in the data center are sorted in ascending order of FP-Score and the hosts are categorized to three levels as in Fig. 4.

TABLE. II. TRANSFORMATIONS FUNCTIONS

Input Layer	No transformation
Hidden Layer	Gaussian Activation function
Output Layer	Linear Activation function

TABLE. III. NEURAL PARAMETERS

N	number of output layer neurons
$\propto_{j0}$	basis to the j output neuron
$\propto_{jk}$	Weight connecting j to k th neuron
$\Phi_k(X^i)$	The response of k th hidden neuron to input X^i modeled as $\Phi_k(X^i) = exp\left(-\frac{\|\|x^i - \mu_k\|\|^2}{\sigma_k{}^2}\right)$ μ_k, σ_k represent the center and width of kth hidden neuron

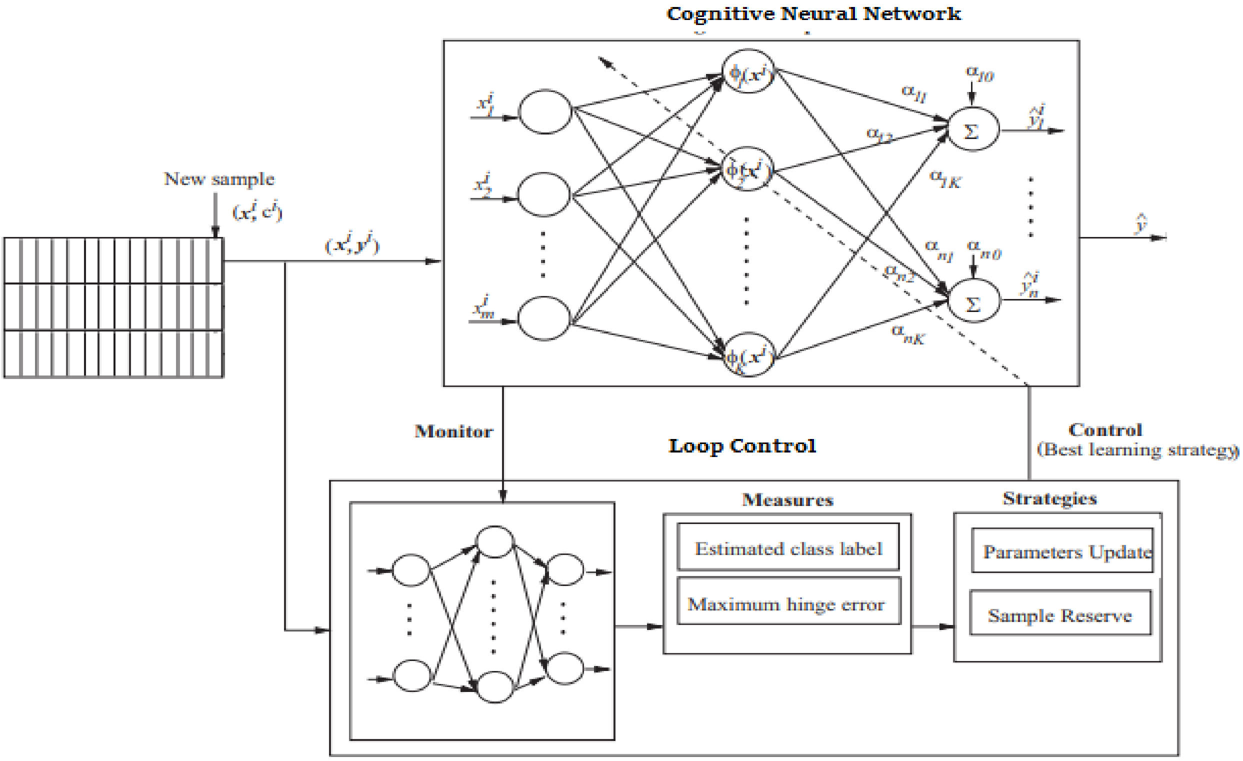


Fig. 2. Deep Learning Classifier.

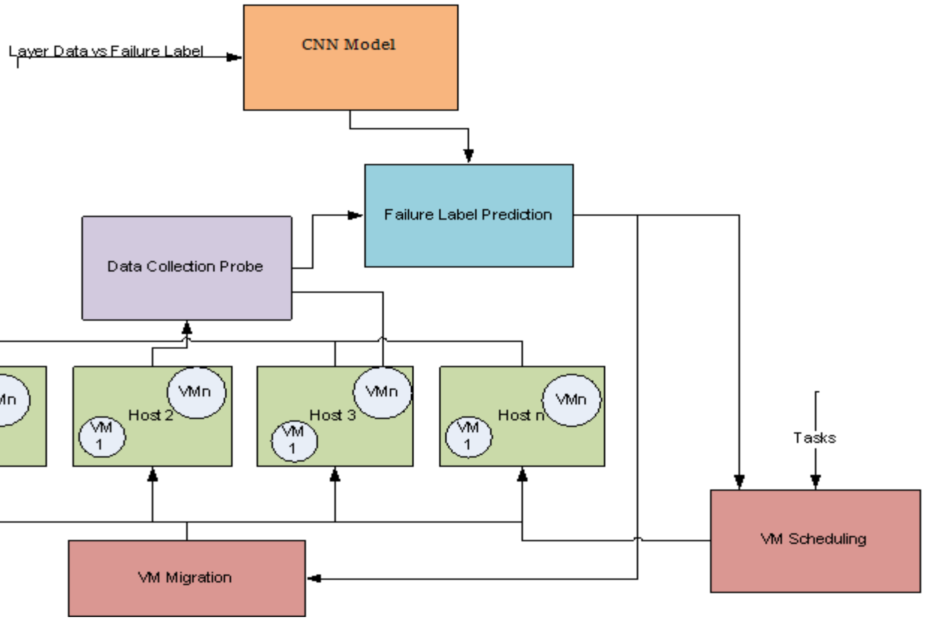


Fig. 3. Solution Architecture.

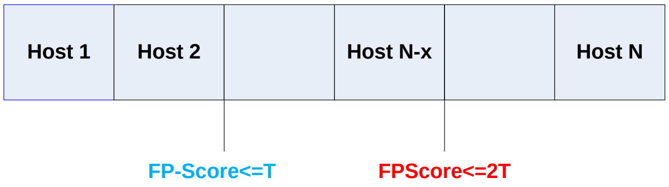


Fig. 4. FP-Score Categorization.

VM's are allocated to host who FP-Score is less than the threshold (T) in the order of first fit. In case VM cannot be allocated to host with FP-Score less than T, the host with score less than 2T is considered with replication in at least 2 hosts. The hosts with FP score greater than T and less than 2T has large chances of failing, so the VM is allocated in redundant manner in this case to handle the failure proactively. The VM are never allocated to host with FP-Score greater than 2T as they has highest changes of failing soon and are in process of migrating their load to reduce their FP-Score.

FP-VM migration policy migrates the VM in host based on their FP-Score. For the host whose FP-Score is greater than 2T, VMs in it are migrated in a progressive manner to other hosts whose score is less than 2T and checked if the FP-Score is reducing , VMs in host are migrated till the FP-Score reaches less than T. During migration, if a favorable host less than T cannot be found, VM is migrated to hosts <=2T and the VM is tried for replication too. If the replication fails, then VM is check pointed. Even in case of VM cannot be migrated due to hosts unavailable, the VM is check pointed. Check pointing is done as the last decision, so that in case of host fails, the VM can be restored till the last check point. Check pointing is a resource consuming task and it is done only as the last resort. For a check pointed VM, if favorable host less than T becomes available later, the check pointing process is immediately abandoned. The pseudo code for VM allocation and migration in the proposed solution is given in Fig. 5 and 6.

```
Algorithm : FP-VM Allocate
Input: VM, T

Sort Hosts based on FP-Score in ascending order
Split Hosts to three ranges ≤T, ≤2T,>2T
R=Allocate VM to Hosts ≤T
If R is success return;
R=Allocate VM to 2 different Host in  T<FPScore≥2T
If R is success return;
Skip the VM allocation till some time
```

Fig. 5. VM Allocation.

```
Algorithm : FP-VM Migrate
Input: VM, T

Sort Hosts based on FP-Score in ascending order
Split Hosts to three ranges ≤T, ≤2T,>2T
R=Migrate VM in host >2T to <T
If R is success return;
R=Migrate VM to Host in  T<FPScore≥2T
If R is success migrate VM to T<FPScore≥2T
   R2== Replicate VM in  T<FPScore≥2T
   If R2== fail , checkpoint VM
If R is fail , check point VM
```

Fig. 6. VM Migration.

IV. Results

Performance evaluation of the proposed solution is done on Cloud-Sim [16] test bed for various simulation configurations. The cloudlet, VM, Host configurations for simulation test bed is imported from Google cluster trace. The performance is measured in terms of

1) Response time
2) Deadline Miss Ratio
3) VMs Allocated

The performance of the proposed solution is compared with QFEC solution proposed in [1].

The response time for completion of task is measured for varied number of tasks and the result is below.

From the results in Fig. 7, it can be seen that the response time is comparatively less in the proposed FP-CNN model. It is due to reduction in the amount of replication and wastage of resources consumed by replicated tasks. Deadline miss ratio is measured for varied number of tasks and the result is below.

From the results in Fig. 8, it can be seen in the proposed system, the deadline miss ratio is less than 20% as against 55% in case of QFEC method. The number of VM's allocated for replication is measured for varying number of tasks and the result is below.

From the results in Fig. 9, The number of VM allocated for replication is less in the proposed method compared to QFEC. The number of instances of intolerance to failure is measured for varied number of tasks and given below.

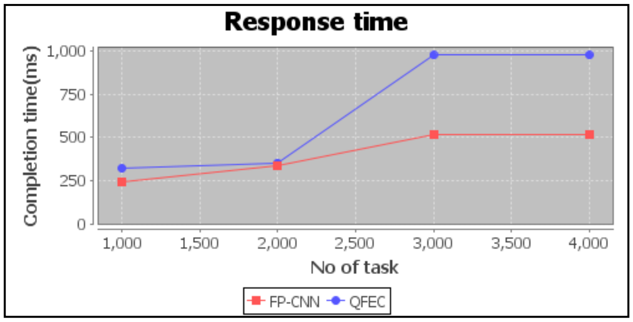


Fig. 7. Response Time.

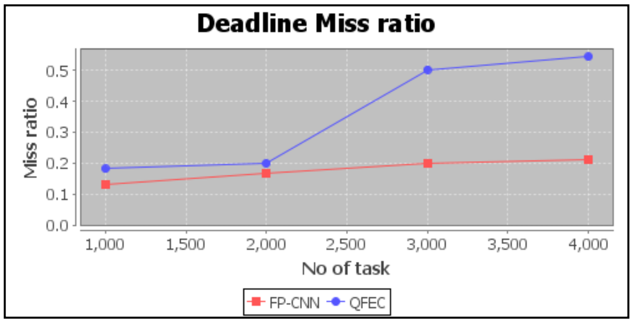


Fig. 8. Deadline Miss Ratio.

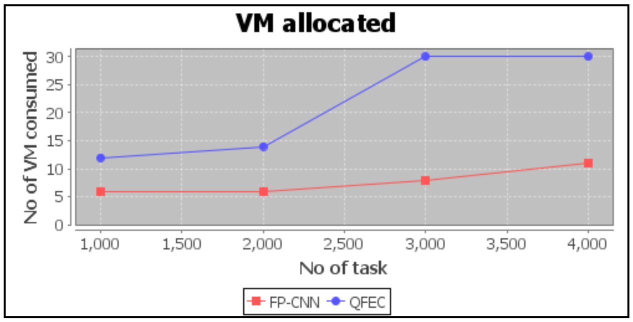


Fig. 9. VM Allocated.

V. Conclusion

A Deep learning based proactive fault tolerance solution called FP-CNN is proposed. Each host in datacenter is assigned to a FP-Score which indicates the failure probability of host. The FP-Score allocation is based on features extracted from various layers and deep learning on those features. Two novel solutions for VM allocation and VM migration is proposed integrating the FP-Score value of hosts. Performance was evaluated in Cloud-Sim platform against Google cluster dataset and from the results, the effectiveness of the proposed solution in providing a proactive fault tolerance is proved.

References

[1] Guoqi Xie,Gang Zeng "Quantitative Fault-Tolerance for Reliable Workflows on Heterogeneous IaaS Cloud",IEEE TRANSACTIONS ON CLOUD COMPUTING 2017.

[2] Jialei Liu,Shangguang Wang,Ao Zhou "Using Proactive Fault-Tolerance Approach to Enhance Cloud Service Reliability",IEEE Transactions on Cloud Computing May 2016.

[3] Z. Zheng, T. Zhou, M. Lyu, and I. King, "Component ranking for faulttolerant cloud applications," IEEE Transactions on Services Computing, vol.5, no.4, pp. 540-550, 2012.

[4] Fan G, Yu H, Chen L, Liu D. Editors. Model based byzantine fault detection technique for cloud computing. 2012 IEEE Asia-Pacific Services Computing Conference (APSCC);Guilin. 2012 Dec 6-8. p. 249–56.

[5] E. AbdElfattah, M. Elkawkagy, and A. El-Sisi, "A reactive fault tolerance approach for cloud computing," in 2017 13th International Computer Engineering Conference (ICENCO), 2017, pp. 190– 194.

[6] M. K. Gokhroo, M. C. Govil, and E. S. Pilli,"Detecting and mitigating faults in cloud computing environment," 3rd IEEE Int. Conf. , 2017.

[7] D. Yuan, D. Park, P. Huang, Y. Liu, M. Lee, Y. Zhou, and S. Savage, "Be Conservative: Enhancing Failure Diagnosis with Proactive Logging," in USENIX OSDI, 2012, pp. 293–306.

[8] S. Kadirvel, J. Ho, and J.AB Fortes, Fault management in map-reduce through early detection of anomalous nodes, in10th International Conference on Autonomic Computing (ICAC 13), California, USA, Jun.2013, pp.235-245.

[9] Y. Tan, X. Gu, and H. Wang,Adaptive system anomaly prediction for large-scale hosting infrastructures,in 29th ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing, Zurich, Switzerland, Jul. 2010, ACM, pp. 173-182.

[10] M. Gabel, R. G. Bachrach, N. Bjorner, and A. Schuster, Latent fault detection in cloud services,Microsoft Research, Tech. Rep. MSR-TR2011-83, 2011.

[11] Q. Zhu, T. Tung, and Q. Xie,Automatic fault diagnosis in cloud infrastructure,in IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom), Bristol, UK, Dec.2013, vol. 1, pp. 467-474.

[12] P. Bodik, M. Goldszmidt, A. Fox, D. Woodard, and H. Andersen, "Fingerprinting the datacenter: automated classification of performance crises," in Proceedings of the 5th European conference on Computer systems (EuroSys2010), Paris, France, 2010, pp. 111–124.

[13] Leonardo Mariani,Cristina Monni,Mauro Pezzé "Localizing Faults in Cloud Systems",IEEE 11th International Conference on Software Testing, Verification and Validation 2018.

[14] Daniel Joseph Dean, Hiep Nguyen, and Xiaohui Gu. "UBL: Unsupervised Behavior Learning for Predicting Performance Anomalies in Virtualized Cloud Systems". In: Proceedings of the International Conference on Autonomic Computing. ICAC '12. ACM, 2012, pp. 191–200.

[15] Z. Li and L. Itti, "Saliency and gist features for target detection in satellite images," IEEE Trans. Image Process., vol. 20, no. 7, pp. 2017–2029, Jul. 2011.

[16] CloudSim:http://www.cloudbus.org/cloudsim/.

Deep Learning Classification of Biomedical Text using Convolutional Neural Network

Rozilawati Dollah[1], Chew Yi Sheng[2], Norhawaniah Zakaria[3], Mohd Shahizan Othman[4], Abd Wahid Rasib[5]
School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia[1, 2, 3, 4]
Program of Geoinformation, Faculty of Built Environment and Surveying,Universiti Teknologi Malaysia[5]
81310 Johor Bahru, Johor, Malaysia

***Abstract*—In this digital era, the document entries have been increasing days by days, causing a situation where the volume of the document entries in overwhelming. This situation has caused people to encounter with problems such as congestion of data, difficulty in searching the intended information or even difficulty in managing the databases, for example, MEDLINE database which stores the documents related to the biomedical field. This research will specify the solution focusing in text classification of the biomedical abstracts. Text classification is the process of organizing documents into predefined classes. A standard text classification framework consists of feature extraction, feature selection and the classification stages. The dataset used in this research is the Ohsumed dataset which is the subset of the MEDLINE database. In this research, there is a total number of 11,566 abstracts selected from the Ohsumed dataset. First of all, feature extraction is performed on the biomedical abstracts and a list of unique features is produced. All the features in this list will be added to the multiword tokenizer lexicon for tokenizing phrases or compound word. After that, the classification of the biomedical texts is conducted using the deep learning network, Convolutional Neural Network which is an approach widely used in many domains such as pattern recognition, classification and so on. The goal of classification is to accurately organize the data into the correct predefined classes. The Convolutional Neural Network has achieved a result of 54.79% average accuracy, 61.00% average precision, 60.00% average recall and 60.50% average F1-score. In short, it is hoped that this research could be beneficial to the text classification area.**

Keywords—Convolutional neural network; biomedical text classification; compound term; Ohsumed dataset

I. Introduction

As a result of the increasing growth rate of the online biomedical text, researchers often encounter a lot of tough challenges. Due to the huge amount of biomedical document entries published online every day, the task of organizing the documents is getting more and more difficult. Eventually, researchers might retrieve irrelevant documents from online sources and they have to waste their precious time to filter and check whether the documents they found is the one they need actually. To deal with all these challenges and problems, classification is one of the effective yet efficient ways. Therefore, in this research, classification of the biomedical text by using deep learning neural network, Convolutional Neural Network will be the focus. By using CNN, researchers can save their precious time in searching for all the intended information.

A. Problem Background

Biomedical literature is papers of scientific research which consists of the convincing idea, theories and results of research done in the medical field. The coverage of biomedical literature or journals are very wide, it could be research about the discovery of new drugs and cure for certain diseases, fresh yet useful information about certain diseases, the discovery of new protein and so on. These papers' main function is to allow communication between the researchers and the other researchers, scientist or even people who may not be trained as a scientist or physician such as students and so forth. Most of the time, biomedical literature are very long and complicated, it might be consisted of up to hundred pages and in these papers, there are also a lot of complex scientific terms which make it more difficult for the researchers to read and find their desired information. With the overwhelming number of biomedical literature available nowadays, researchers are facing a lot of difficulties when they wanted to retrieve their desired information in their field of study [1].

Text classification is an important component in many applications, for instance, web searching, information filtering and sentiment analysis [2][3]. By doing text classification, we are assigning predefined categories to the free-text documents [4]. It is also can be known as text categorization or document categorization. The classification of text is mainly done by using machine learning algorithms such as feature engineering, feature selection and so forth. However, text classification using machine learning approaches might have data sparsity problem [5]. To deal with this problem, we apply deep learning principle, Convolutional Neural Network (CNN) instead of the machine learning approach.

Recently, deep learning approaches have also been used to do text classification. Convolutional Neural Network is one of the deep neural networks and it is believed that this neural network can solve the data sparsity problem [6]. It is very useful in extracting information from raw signals, ranging from computer vision applications to speech recognition and so forth [7]. In this research, we will focus on using Convolutional Neural Network to classify biomedical text abstracts and to measure the effectiveness of using Convolutional Neural Network in text classification.

II. Related Works

Deep learning has gained attractions from may researchers recently as the deep learning approach in classification can produce good results which can be compared to the machine learning approach. A well known deep learning-based approach, Convolutional Neural Network is proposed to perform granite rocks classification [8]. The results obtained from the experiment showed the performance of some of the networks would be better when they are applied together. Convolutional Neural Network in image classification on the different type of datasets such as remote sensing data of aerial images and scene images from SUN database [9]. The experimental results based on the graphical representation and quality metrics showed that CNN can produce a fairly good result for all the tested datasets. They concluded that a low Mean Squared Error (MSE), high classification accuracy and shorter training time can be achieved by using enough number of epochs, the number of iterations.

CNN was also applied to the biomedical domain text classification to identify the hallmarks of cancer associated with publication abstracts [10]. The data that had been used in this research is the text document from the online source corpus. The experimental results showed that a competitive performance with the state-of-the-art SVM-based classifier can be achieved. A simple CNN with only one convolution layer can perform notably well and an unsupervised pre-training of the work vectors is very essential in deep learning for natural language processing [11].

A recurrent convolutional neural network to perform text classification [5]. The proposed model is outperforming the traditional recurrent neural network and convolutional neural network as the contextual information can be captured with the use of recurrent structure and then the representation of the text is done using the convolutional neural network.

III. Methodology

This research could be divided into five phases, Ohsumed dataset collection phase, text pre-processing phase which involved feature extraction, text classification phase, deep learning architecture phase as well as evaluation and validation phase. The details of each phase will be discussed as follow:

A. *Ohsumed Dataset Collection Phase*

The Ohsumed datasets might contain different levels from the first level until the fourth level and this research will focus on 11,566 abstracts of biomedical journal which are from first and second levels only. All the categories and the number of abstracts for each category used in this research is stated in Table I.

TABLE. I. List of Selected Categories with Document Numbers

Category Name	**Number of Documents**
Arrhythmia	1,173
Coronary Disease	4,235
Heart Arrest	513
Heart Defects, Congenital	718
Heart Failure, Congestive	1,295
Heart Valve Diseases	335
Myocardial Diseases	355
Myocardial Infarction	2,942
TOTAL ABSTRACTS USED: 11,566 ABSTRACTS	

B. *Text Pre-processing Phase*

Text pre-processing is an important process to extract the important and informative features before we can classify the biomedical text [12]. In this phase, feature extraction is done to retrieve the most relevant information from the training set and represent that information in a lower dimensionality space. Feature extraction is performed by using the GENIA tagger tool, a tool which is specially designed to extract information from biomedical text. Fig. 1 shows the output produced by the GENIA tagger tools, all words in the text are labeled with post-of-speech (POS) tags, chunk tags and named entity tags. The outputs produced by GENIA tagger tools will be processed again by selecting only the noun phrases which all the phrases are labeled with NP tag. After that, the stop words, general terms and duplicate terms will be removed from the text as it might be able to influence the result of classification afterward. All the remaining features will be grouped to form a vocabulary with the total number of 22,176 features and all these features will be added into the multiword tokenizer lexicon to be used in the process later.

C. *Deep Learning Architecture Phase*

Generally, deep learning text classification model architecture used in this research consists of several components and all these components are connected sequentially. For instance, the components in the architecture included the input, biomedical abstracts, word embedding layer, deep network which is made up of convolutional layers and max-pooling layer, fully connected layer and the output which is the classification result. The deep learning text classification model architecture used in this research is shown in Fig. 1.

The input which is the biomedical abstracts will go through the process of tokenizing using the multiword tokenizer instead of the single word tokenizer. The reason why multiword tokenizer is used instead of the single word tokenizer is because the dataset used in this research is all biomedical abstracts which contain many biomedical terms, all these biomedical terms mostly are compound terms which are built by combining two or more single term and these compound terms carry different meaning with the single term.

Next, the word embedding layer must be set up before the pre-processed text input passes through. The purpose of this layer is to transform all the words in the text that have the same or similar meaning to have a similar representation in the form of a vector, it is a good technique which can be used to acquire continuous low-dimensional vector space representation of words [13]. Keras, a deep learning framework used in this research provides the embedding layer to handle this word embedding; it stores a lookup table for mapping between the words in the biomedical abstracts and the dense vector representations. In this research, a pre-trained BioASQ word vector is used [14].

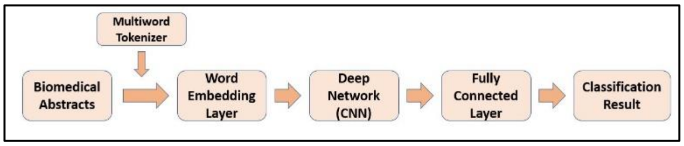


Fig. 1. Deep Learning Text Classification Model Architecture.

The sequence of embedding vectors obtained from the previous process will be converted into a compressed representation with all the information in the sequence of words in the text captured completely and afterward the stack of convolutional layer and max-pooling layer take it as the input. First, the convolutional layer consists of trainable kernels which also known as filters which is functioned to detect any specific features in the input. They will slide or convolve across the one-dimensional input and produce an activation map. A set of activation maps is produced from the different convolution process which detects different features and passed to the max-pooling layer.

The activation function used in the proposed model is the Rectified Linear Unit (ReLU). ReLU function does not activate all neurons at the same time and hence it is very efficient in term of computational time. Next, the max-pooling layer will take the transformed output from the convolutional layer as input and this layer functions to reduce the computation complexity and the spatial dimension without dumping the momentous information. In this research, the momentous information mentioned before this is the significant features. The output from the max-pooling layer will be taken as the input of the fully connected layer.

Next, the fully connected layer is where the classification performed based on the features extracted from the stack of convolutional layers and max-pooling layers. In this research, softmax function with categorical_crossentropy loss function is used as we are solving a multi-class classification problem. Softmax function here functions to apply a transformation to the output obtained from the previous layers so that the final output can be interpreted as a probability vector for each class or class scores.

During the training process of the model, the cross-validation method is used to reduce the problems like model overfitting and give an insight on how well the model can generalize to each independent dataset. In this research, stratified k-fold cross-validation with *k* equals to 10 are used to evaluate the performance of the model as 10 folds is believed to have the smallest mean squared error as well as variance in the estimation of prediction errors [15]. Stratified 10 folds validation split the entire dataset into 10 parts and it will take one part of the dataset to test the model while the other parts of the dataset will be used to train the model. Besides, it takes at least *m* instances from each of the classes for the training process at each fold to prevent a situation where the data from the certain classes are over-represented. Cross-validation is important to ensure that every single part of the dataset is used is to train the model. The parameters used for the proposed model is shown in Table II.

D. Performance Evaluation and Validation Phase

During the training process, the dataset is split into the training set and validation set. In this research, the training set consists of 80% of the whole dataset while the validation set consists of 20% of the whole dataset. The training set is purposely used to train the model while the validation set is used to evaluate the model's performance. Metrics on the training set like the training loss and training accuracy allow us to know the progress of the model in terms of training while the metrics on the validation set like the validation accuracy and validation loss allow us to measure the quality of the model in terms of the capability to make prediction based on the new data.

TABLE. II. DESCRIPTION AND VALUE OF USED PARAMETERS FOR THE PROPOSED MODEL

Parameter	**Description**	**Value**
Epoch	Number of the forward and backward pass of all the training samples	10
Filter size	The dimension of the filter in the ConvNet	128
Kernel size	Size of the convolutional filter	3
Batch size	The total number of datasets that will be propagated through the network	128

Generally, a good model should have perfect fitting where the training loss is roughly the same as the validation loss. To reduce the effect of overfitting, dropout can be inserted after each max-pooling layer. This can also decrease the training time and result in better performance. In this research, there are other evaluation methods such as precision, recall and F1-measure also have been used to evaluate the proposed model. In this research, these three evaluation methods are performed by using the scikit-learn classification report. The formula to calculate the precision, recall and F1-measure are as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F1- score} = 2 * \frac{precision*recall}{precision+recall} \quad (3)$$

Where TP is true positive, FP is false positive and FN is false negative.

IV. EXPERIMENTAL RESULTS

There are many sets of experiments have been carried out to evaluate the performance of the proposed model. All these experiments are evaluated based on the classification report, the values of precision, recall, and F1- score. The manipulated methods used to test the proposed model included the utilization of single word tokenizer or the multiword tokenizer as well as the number of sets of convolutional and max-pooling layers used in the Convolutional Neural Network architecture.

The sequence of embedding vectors obtained from the previous process will be converted into a compressed representation with all the information in the sequence of words in the text captured completely and afterward the stack of convolutional layer and max-pooling layer take it as the input. First, the convolutional layer consists of trainable kernels which also known as filters which are functioned to detect any specific features in the input. They will slide or convolve across the one-dimensional input and produce an activation map. A set of activation maps is produced from the different convolution process which detects different features and passed to the max-pooling layer.

Overall, the performance of the proposed model using the multiword tokenizer is decreasing as the number of sets of the convolution and the max-pooling layer is increasing in terms of the precision, recall as well as the F1-measure. This is because one set of the convolution and max-pooling layer is enough to extract features from the biomedical text and carry out the classification of the biomedical text. As the number of sets of convolution and max-pooling layer increasing, the model tends to extract more features which are irrelevant to be used in the classification task. The performance measure for the experiment using multiword tokenizer with one set of the convolutional and the max-pooling layer is shown in Fig. 1.

The performance of the proposed model using the single word tokenizer has the same trend as the one using multiword tokenizer which is the performance of the model is declining as the number of sets of layers is increasing. However, the performance of the model using the single word tokenizer is better compared to the model that used multiword tokenizer. The performance measure for the experiment using multiword tokenizer with one set of the convolutional and max-pooling layer is shown in Fig. 2.

Looking deeper into the performance measure obtained for both sets of experiments, we can see that the category with the high number of documents is having a better performance compared to the category with the low number of documents. For instance, categories like coronary disease and myocardial infarction which consists of 4,235 documents and 2,942 documents respectively, they have achieved better overall performance compared to other categories. On the other hand, categories like Heart Valve Diseases and Myocardial Diseases that consists of 335 and 355 documents which considered as a very low number of documents, they have the worst precision, recall and F1-score in all three sets of the experiment conducted which are 0 for precision, recall and F1-measure. The model performance on categories like Heart Valve Diseases is completely zero in terms of precision, recall and F1-score. This is because, in each category, there are different biomedical terms, as this category has a smaller size compared to the other categories, the features that can be used to classify the biomedical text will be less as well (see Fig. 3).

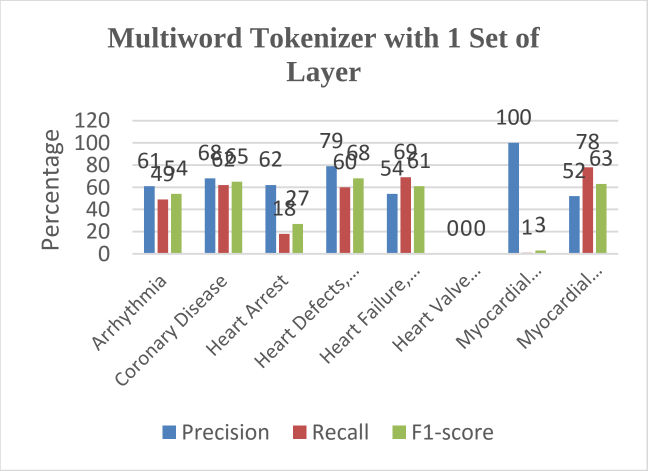


Fig. 2. Result of the Experiment using Multiword Tokenizer with 1 Set of Convolution and Max-Pooling Layer.

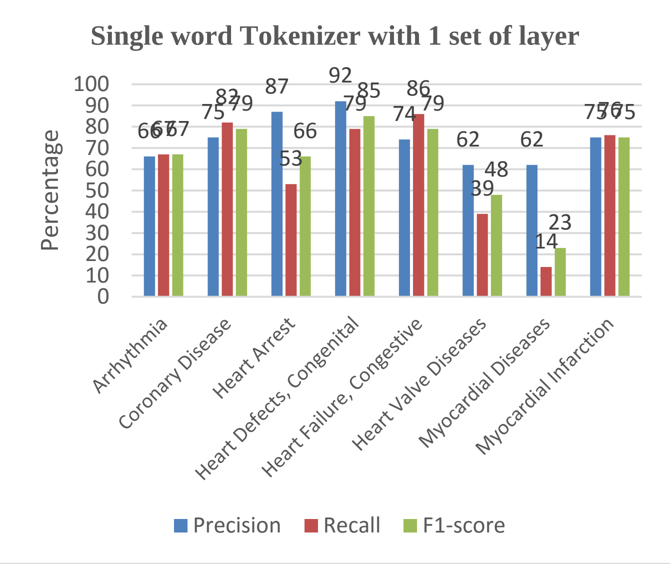


Fig. 3. Result of the Experiment using Single Word Tokenizer with 1 Set of Convolution and Max-Pooling Layer.

TABLE. III. COMPARISON BETWEEN THE RESULT OF THE EXPERIMENT USING MULTIWORD TOKENIZER AND SINGLE WORD TOKENIZER

	Average Accuracy (%)	Average Precision (%)	Average Recall (%)	Average F1-score (%)
Using multiword tokenizer	54.79	61.00	60.00	60.50
Using single word tokenizer	70.64	75.00	75.00	75.00

Fig. 4 and Table III demonstrate the comparison between the experimental result using the multiword and single word tokenizer. In this comparison, both experiments are using the same architecture, one set of convolution and max-pooling layer since this is the most ideal architecture that achieved the best performance among the others. The model with the use of single word tokenizer is outperforming the model with the used of the multiword tokenizer. This is because the single word tokenizer tokenizes each word in the text into single tokens which carry less or no meaning to pass through the word embedding layer. Unlike what is done by the multiword tokenizer which tokenizes the words in the text according to the words added into the lexicon at the early stage. The outcome would be a list of tokens which included single terms as well as the compound terms. This is important as the dataset used in this research is biomedical text which consists of a lot of biomedical terms that should be taken in consideration when doing text classification.

Overall, the result of experiment II is best among the three sets of experiments as shown in Fig. 5. Although the average precision, recall and F1-score did not show a very satisfying value, but if we compare the average precision and average recall, we can see that in Experiment II, the average precision is 69% and the average recall is 67% has only a difference of

2% which indicates that almost all the documents retrieved are relevant. This is because the dataset B has more documents in each category and hence there will be more documents involved in the training process which in turn give a better performance. Experiment III gives the worst result among the three sets of experiments; it achieved 73% for average precision and 60% for average recall. This indicates that in Experiment III, the number of documents retrieved is less relevant as there is a difference of 13% between the value of average precision and average recall. This scenario is caused by a small number of documents involved in the training process. In conclusion, to have a good deep learning model performance, the size of the dataset used must be large enough.

Fig. 6 illustrates the comparison of training accuracy and validation accuracy for the model using multiword tokenizer and one set of convolution and max-pooling layer as the model architecture and the number of epochs in this experiment is set to 10 epochs. We can observe that both the training accuracy and validation accuracy have the same trend of increasing. Overfitting is a scenario where the model tends to memorize the data instead of learning it. The degree of overfitting is indicated by the gap between the training accuracy line and the validation accuracy line, smaller gap means less overfit and vice versa. In this research, the degree of overfitting is minimized by adding dropout after the convolution and max-pooling layer. The degree of overfitting can also be known by comparing the training loss and validation loss. In a model with the perfect fit, the training loss should always lower and roughly the same with the validation loss. In this case, the difference between the training loss and the validation loss is not more than 0.2 which indicates that the added dropouts have successfully reduced the degree of overfitting.

The results of this research is compared to the result conducted by Hughes *et al.* [16] and it is shown in Fig. 7, we can see that the methods used in this research which is the combination of CNN, Multiword Tokenizer and Word2vec is outperformed compared to the (BOW + LogR) methods, this might be caused by the features used for classification in our proposed method are more informative than the features used in (BOW + LogR) method. On the other hand, the proposed method in our research is less perform compared to (CNN + Word2vec), because using different tokenizer will result in a different list of features. In their research, they use only the single word tokenizer and the features used for the classification might be less meaningful compared to the features used in our research, for instance, features “lung” and “cancer” are less informative than feature “lung cancer” and hence resulting the better performance in terms of accuracy, 68% compared to our proposed method which achieved only 54%. Hence, it can be concluded that a CNN-based approach with the use of multiword tokenizer can be used to conduct biomedical text classification and produce a better performance.

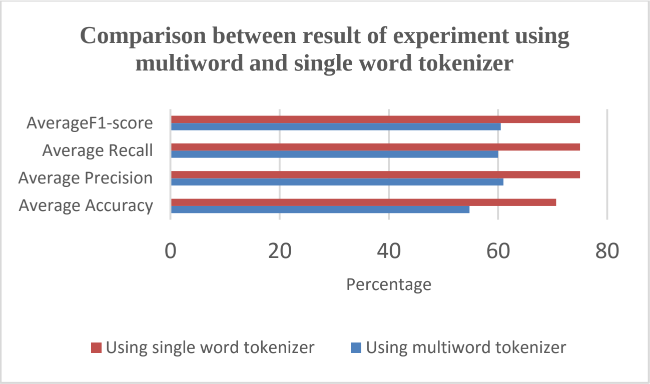


Fig. 4. Comparison between the Result of an Experiment using Multiword and Single Word Tokenizer.

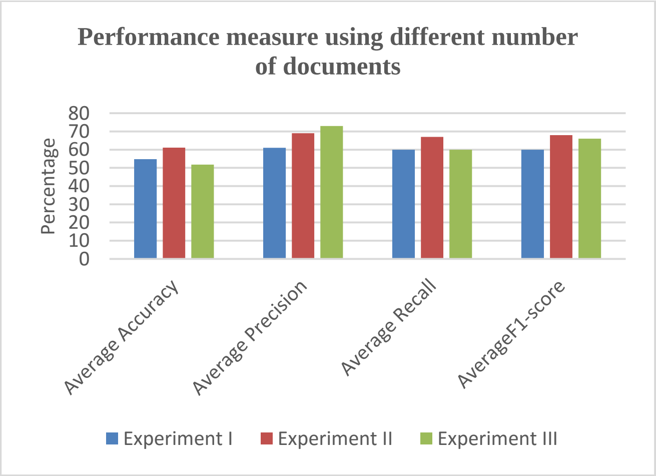


Fig. 5. Performance Measure using different Number of Documents.

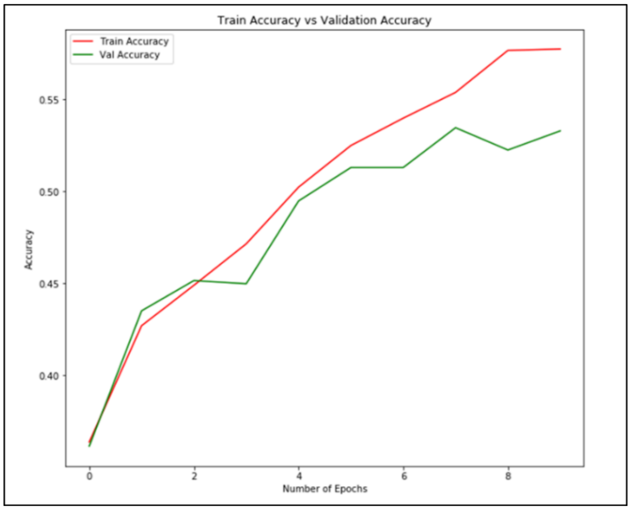


Fig. 6. Comparison of Training Accuracy and Validation Accuracy.

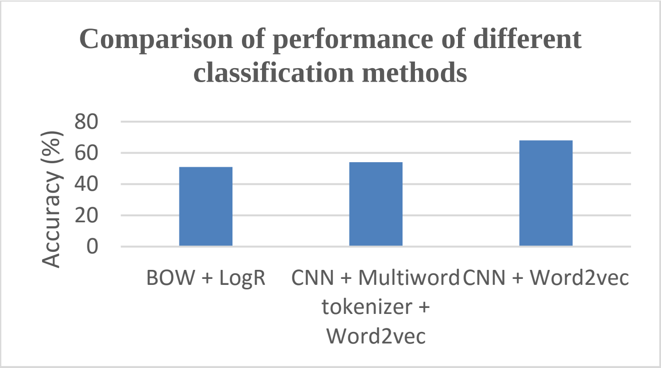


Fig. 7. Comparison of Performance of different Classification Methods.

V. Conclusion

In this study, Convolutional Neural Network was used to perform classification of biomedical text and a result of average accuracy of 54.79%, average precision of 61.00%, average recall of 60.00% and average F1-score of 60.50% were obtained. It did take consideration of the biomedical terms in the biomedical text by using multiword tokenizer, biomedical terms which are mostly made up of two or more terms were tokenized into compound tokens and used for the CNN to perform text classification. All and all, the proposed method can increase the recall percentage, in other words, increase the number of documents being retrieved and classified correctly and indeed it is a good approach to be used in the classification of biomedical text.

VI. Future Works

Firstly, a new word vector could be retrained by using all the biomedical text from the entire online biomedical text repository instead of just using the pre-trained BioASQ word vector which involves only biomedical text from PubMed. This is to ensure that the single term tokens or compound term tokens both can be assigned with meaningful vectors when they are passing through the word embedding layer. Next, a dataset which is greater in size could be used. This is because a deep learning neural network needs a lot of data in the training process. In other words, it needs a lot of data for the learning purpose. All in all, CNN is really an effective yet efficient approach to do classification tasks.

Acknowledgment

This study is supported by the Fundamental Research Grant Scheme (FRGS) under the Vote No. 4F559 that sponsored by Ministry of Higher Education (MOHE), Research University Grant Scheme (RUG) under the Vote No. 13J94 and UTM Industry/International Incentive Grant (IIIG) under the Vote No. 4C255. The authors are greatly obliged to Universiti Teknologi Malaysia (UTM) and Soft Computing Research Group (SCRG) for the support and motivation.

References

[1] M. Pavlinek, and V. Podgorelec, “Text classification method based on self-training and LDA topic models,” Expert Systems with Applications, vol. 80, pp. 83–93, September 2017.

[2] C. C. Aggarwal, and C. Zhai, “A Survey of Text Clustering Algorithms,” Mining Text Data, Springer, pp. 77-128, January 2012.

[3] M. Labani, P. Moradi, F. Ahmadizar, and M. Jalili, “A novel multivariate filter method for feature selection in text classification problems,” Engineering Applications of Artificial Intelligence, vol. 70, pp. 25-37, April 2018.

[4] M. M. Mirończuk, and J. Protasiewicz, “A recent overview of the state-of-the-art elements of text classification,” Expert Systems with Applications, vol. 106, pp. 36–54, September 2018.

[5] S. Lai, L. Xu, K. Liu, and J. Zhao, (2015). Recurrent Convolutional Neural Networks for Text Classification. Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 2267–2273.

[6] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, “Recent advances in convolutional neural networks,” Pattern Recognition, vol. 77, pp. 354–377, May 2018.

[7] X. He, and W. Zhang, “Emotion recognition by assisted learning with convolutional neural networks,” Neurocomputing, vol. 291, pp. 187–194, May 2018.

[8] A. Ferreira, and G. Giraldi, “Convolutional Neural Network Approaches to Granite Tiles Classification Convolutional Neural Network approaches to granite tiles classification,” Expert Systems With Applications, vol. 84, pp. 1–11, October 2017.

[9] D. Jaswal, V. Sowmya, and K. P. Soman, “Image Classification Using Convolutional Neural Networks,” International Journal of Advancements in Research & Technology, vol. 3, Issue 6, pp. 1661–1668, June 2014.

[10] S. Baker, A. Korhonen, and S. Pyysalo, “Cancer Hallmark Text Classification Using Convolutional Neural Networks,” in Proceeding of the Fifth Workshop of Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), pp. 1–9, December 2016.

[11] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751, October 2014.

[12] V. Gurusamy, and S. Kannan, “Preprocessing Techniques for Text Mining,” in Conference Paper, India, October 2014.

[13] C. Y. Liang, W. Jin, K. R. Lai, and Z. Xuejie, “Refining Word Embeddings for Sentiment Analysis,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 534-539, September 2017.

[14] I. Pavlopoulos, A. Kosmopoulos, and I. Androutsopoulos, “Continuous Space Word Vectors Obtained by Applying Word2Vec to Abstracts of Biomedical Articles,” Word Journal Of The International Linguistic Association, pp. 1-4. March 2014.

[15] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 2, pp. 1137–1143, August 1995.

[16] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, “Medical Text Classification using Convolutional Neural Networks,” Studies in Health Technology and Informatics 235, pp. 246–250, April 2017.

Acquisition and Classification System of EMG Signals for Interpreting the Alphabet of the Sign Language

Alvarado-Diaz Witman[1], Meneses-Claudio Brian[2], Fiorella Flores-Medina[3]
Patricia Condori[4], Natalia I. Vargas-Cuentas[5], Avid Roman-Gonzalez[6]
Image Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades, Lima, Peru

***Abstract*—Taking into account that in Peru, there is an increase in people with difficulties in speaking or communicating. According to the National Institute of Statistics and Informatics of Peru (INEI for its acronym in Spanish), around 80000 people use the gesturing language. For this reason, this research proposes to use the electromyography (EMG) signals to detect the hand movement and identify the alphabet of the sign language to provide essential communication to people who need it. The idea is to classify the signals and recognize the letters of the Spanish alphabet, interpreted in the Peruvian sign language. The results show the classification of the 27 letters of the alphabet with a general success rate of 93.9%.**

Keywords—Electromyography; EMG; sign language; essential communication; recognize the letters

I. Introduction

In the national report of the sociodemographic profile [1], carried out by the National Institute of Statistics and Informatics (INEI for its acronym in Spanish), it mentions that in Peru 10.4% of the population suffers from a disability, of which 3.1% have difficulty to speak or communicate. Based on this problem, the present work proposes the use of electroencephalography (EMG) signals in order to provide essential communication for people with difficulties to speak, for this purpose the classification of the signals, produced in the muscles is performed by performing gestures that denote the letters of the alphabet.

The Ministry of Education of Perú, through the General Direction of the Basic Special Education (DIGEBE for its acronym in Spanish) makes available the document "Peruvian Sign Language" [2], in which it is mentioned that sign language is a communication system which is perceived through sight and requires the use of the hands as active articulators, the use of space as a place of articulation (phonological structure) and as a temporary reference.

To obtain data, the Myoware system was used as a basis [3] which is an electronic system to obtain the electrical activity produced by the muscles, they are commonly used in prostheses, robotics and much more; the Myoware detects the electrical activity of the muscles, and then convert it into a variable voltage that can be read in the analog input pin of any microcontroller, in our case, it will use the Arduino UNO board.

In [4] the use of the electromyogram (EMG) data provided by the Myo bracelet is evaluated as characteristics for the classification of 20 stationary letter gestures of the Brazilian Sign Language alphabet. The classification was done using Support Vector Machines (SVM). The results obtained show that it is possible to identify the gestures.

In [5], they present an interface to capture facial muscle data. This system is based on the recognition of syllables of the Spanish language. The signals of each articulated syllable (with silent voice) were transformed into a vector of characteristics that are used to train a classification algorithm. Experimental results demonstrated effectiveness in three subjects, providing an average classification rate of almost 70% for 30 classes.

In [6] a TDM (Time derivative moments) feature set extraction is proposed to improve the performance of EMG-PR (Electromyography Pattern Recognition) in the classification of upper limb movement. They use a standard database to examine the proposed feature set. Finally, 97.6% of effectiveness was obtained in the classification of 8 different kinds of movements.

In [7] an American sign language recognition system is proposed using Electromyography; its objective was to recognize the letters of the sign language alphabet and allow users to spell words and sentences. The signals were acquired from the right forearm for 27 gestures. Methods extraction such as Time domain, frequency domain (band power), power spectral density (band power), and average power features were used. Also, as a classification method, Support Vector Machine (SVM) and Ensemble Learning algorithm were used; obtaining a percentage of effectiveness of 80%.

The main objective of this research is to provide a basic communication between people who use sign language to communicate, this document presents the primary results obtained; which demonstrate the possibility of using our methodology as a system to interpret sign language, it works continue

Section II presents the methodology that has been followed for the research work. In Section III are the preliminary results obtained, and finally, in Sections IV and V are the respective discussions and conclusions.

II. Methodology

there are different methodologies for interpreting sign language, however the methodology presented below is used due to the ease which the data can be obtained, and because of the experience in the treatment of data. themselves; Finally, in conclusión is an agile and practical methodology.

For the present research work, the methodology to follow is schematized in the block diagram shown in Fig. 1.

A. Data Acquisition

For the data acquisition, the system described in [2] in which the Myoware device is used will be used, this device will be connected to an Arduino Uno board, which will acquire the signals and send them by serial port of the computer. In Fig. 2, one can see the mentioned system, in addition to the acquired signals, by representing with the left hand a random position to check the operation of the system.

A graphical interface is created in Matlab (Fig. 3), in which data will be taken, in addition to performing the signal processing, which will be explained in the next section. The data is saved in files with extension ".mat", which is an extension of Matlab; these files contain the data corresponding to the four channels that the system has.

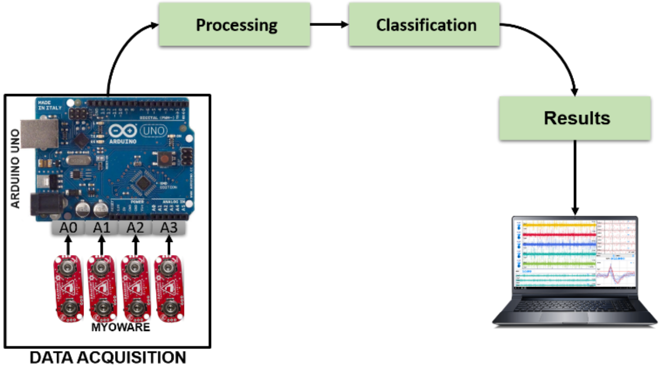


Fig. 1. Block Diagram of the Methodology.

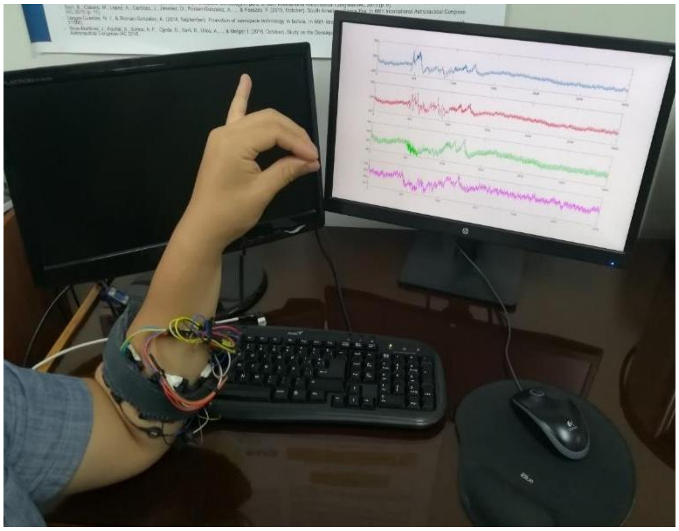


Fig. 2. Data Acquisition.

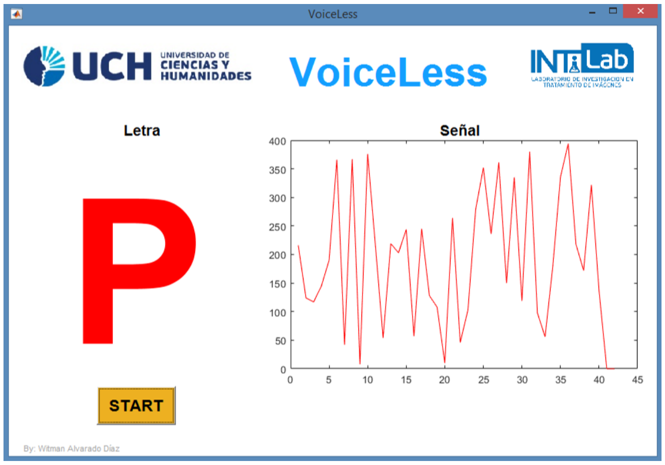


Fig. 3. VoiceLess Graphic Interface.

B. Processing and Classification

For the processing and classification stage, the algorithm of Fig. 4 is used. The classification stage is already included, in the external function, in addition to a prediction stage.

The diagram of the external function can be seen in Fig. 5, it can observe the stage of characteristics extraction, which consists of 8 characteristics of which 4 are from the domain of frequency and the other 4 in the domain of time, the proposed characteristics are: The maximum value, average data, variance, and standard deviation; then these characteristics are grouped and classified as well used to predict the letter made through sign language.

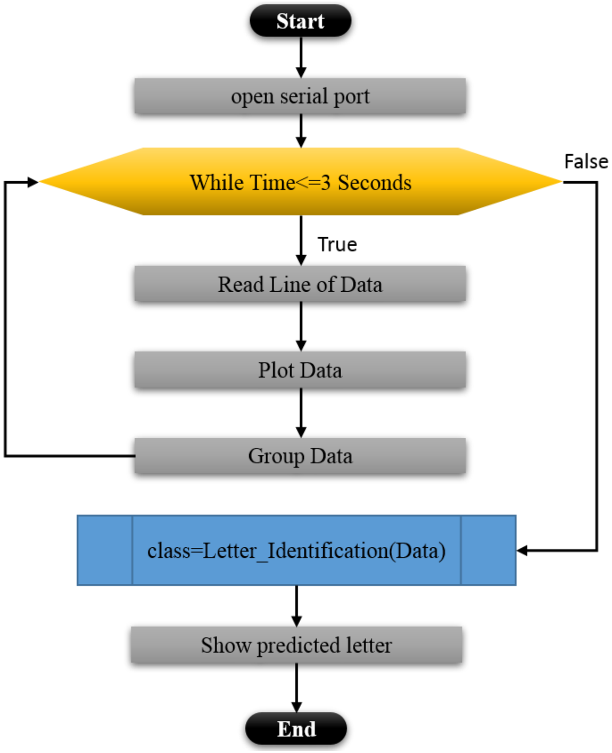


Fig. 4. Flow Diagram of the VoiceLess Graphical Interface.

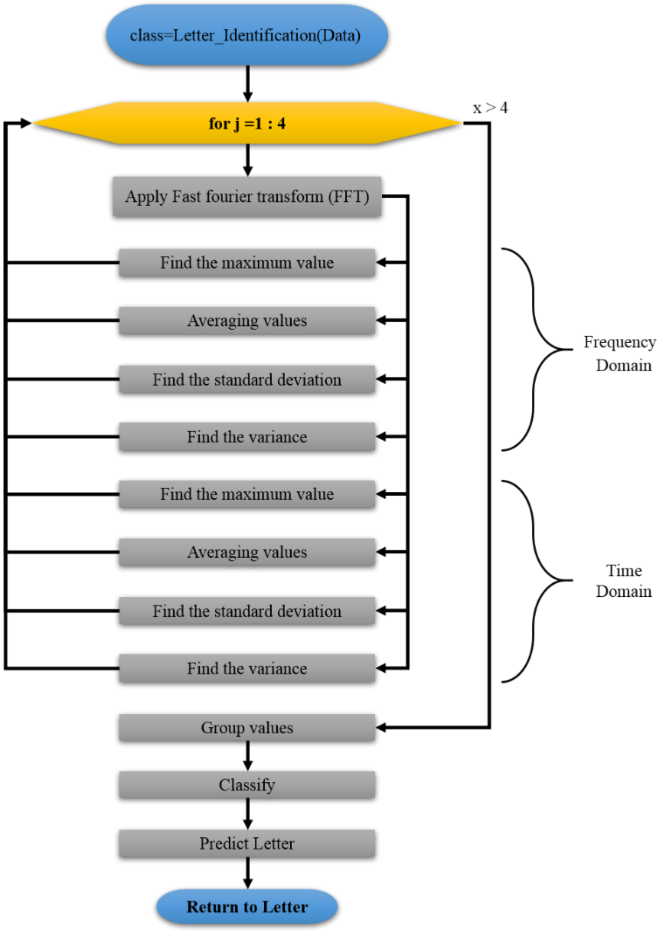


Fig. 5. Scheme of Data Processing.

It has to take in mind that it will work with signals with an amplitude between 0 to 10 mV and a frequency of 10 to 100Hz [8], [9], [10] and [11], [2], and that each data channel is extracted, then finding the fast Fourier transform; next, the above-mentioned characteristics are calculated.

III. Results

Taking in mind that it will first analyze the data that was taken from the left forearm, with the methodology proposed in [2]; Sampling is done as follows: the first 3 seconds no movement is made, the 3 seconds after the representation of a letter of the Peruvian Sign Alphabet is performed, finally in the last 3 seconds no movement is made; It is worth mentioning that samples are taken from a healthy person. Five samples are taken for each of the 27 letters of the alphabet.

In Fig. 6(a), it can see the signals obtained from the representation of the letter “H”, corresponding to channels 1,2,3 and 4; in Fig. 6(b), it can see some of the characteristics obtained from the fourth channel; it can see that there are apparent differences between one and other signals.

The characteristics obtained from all the data were introduced in the Classification Learner application, which can be seen in Fig. 7, which shows the graph of the characteristics of the 27 categories entered.

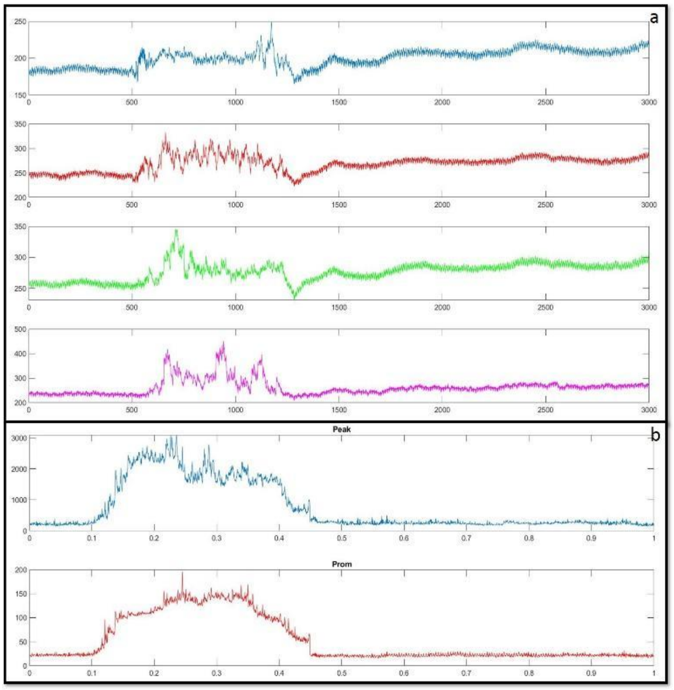


Fig. 6. Raw Signals in Each Data Channel. (b) Graph of the Characteristics of the Data on Channel 4.

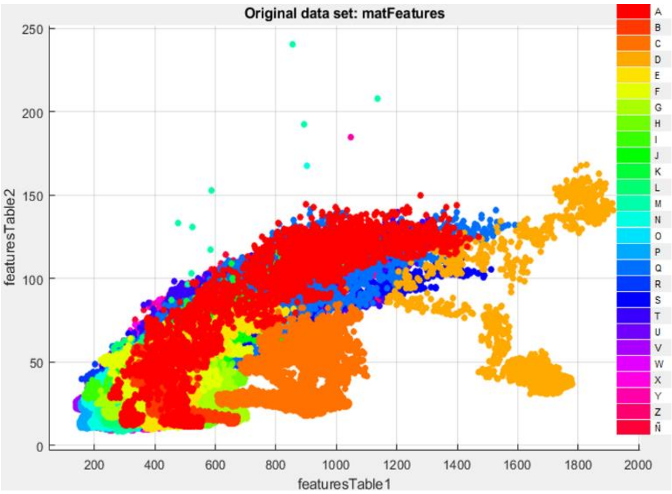


Fig. 7. Characteristics Plotted in the Classification Learner Application.

The same application generates a general rate of some 93.9% and also provides with the necessary data for the construction of a confusion matrix, from which the results shown in Table I are extracted.

From this table, it randomly choose 4 letters “A”, “P”, “U” and “Z” and submit it to the algorithms described in the methodology section, in order to check if the predictions of the letters are made correctly, for which, with the graphical interface, 32 repetitions of each chosen letter are taken, with which the results of Table II were obtained.

As we can see in Table II, ten repetitions were made to the letter “P” was the most successful, followed by the letter “A”, “U” and finally “Z”.

TABLE. I. TABLE OF ACCURACY

Class	**True class**		**Predicted class**	
	True Positive Rate	**False Negative Rate**	**Positive Predictive Value**	**False Discovery Rate**
A	98.8%	1.2%	98.5%	1.5%
B	99.5%	0.5%	99.6%	0.4%
C	100.0%	0%	100.0%	0%
D	100.0%	0%	100.0%	0%
E	98.8%	1.2%	99.1%	0.9%
F	97.5%	2.5%	96.8%	3.2%
G	96.7%	3.3%	98.1%	1.9%
H	97.8%	2.2%	97.9%	2.1%
I	98.9%	1.1%	98.7%	1.3%
J	99.8%	0.2%	97.6%	2.4%
K	99.6%	0.4%	99.3%	0.7%
L	91.7%	8.3%	94.3%	5.7%
M	97.9%	2.1%	97.3%	2.7%
N	93.3%	6.7%	94.5%	5.5%
O	96.9%	3.1%	96.3%	3.7%
P	96.8%	3.2%	97.5%	2.5%
Q	91.8%	8.2%	90.4%	9.6%
R	85.8%	14.2%	86.4%	13.6%
S	87.6%	12.4%	88.9%	11.1%
T	88.8%	11.2%	89.7%	10.3%
U	83.0%	17.0%	89.0%	11.0%
V	91.5%	8.5%	89.5%	10.5%
W	88.1%	11.9%	90.5%	9.5%
X	93.4%	6.6%	92.6%	7.4%
Y	90.5%	9.5%	90.3%	9.7%
Z	94.1%	5.9%	90.2%	9.8%
Ñ	89.0%	11.0%	92.4%	7.6%

TABLE. II. PREDICTED ACCURACY

WORDS	**TRUE**	**FALSE**
A	7	3
P	9	1
U	6	4
Z	5	5

IV. DISCUSSION AND CONCLUSIONS

In [4] for the classification of 20 stationary letter gestures of the Brazilian sign language alphabet, a minimum efficiency of 96.01% was obtained. In the consultation [5], in recognition of Spanish language syllables provides an average classification rate of almost 70% for 30 classes. In the consultation [6] in the classification of the movement of the upper limb through various gestures, 97.6% effectiveness is obtained in the classification of 8 classes. In the consultation [7] in recognition of the letters of the American sign language alphabet, an 80% effectiveness was obtained. In none of the research works consulted was the test carried out if the proposed classifier correctly predicts the letters of the alphabet.

In conclusion, the classification of EMG signals has been achieved. On this research paper, using 4 data channels have reached an overall percentage of correct answers of 93.9% so, to improve the rate of predictions and repeatability of our system, it is necessary to add more samples until a moderate prediction is achieved. As a future work, it is proposed to make 128 samples per letter in addition to improving the hardware and communication with the PC, in order to have a higher sampling frequency of signals, which will show more accurate data of the movements made with hand, it is expected to achieve the objective of VoiceLess, t's to provide essential communication to people are deficient so to communicate.

REFERENCES

[1] INEI, "Perfil Sociodemográfico, Informe Nacional," 2018.

[2] A.-D. Witman, M.-C. Brian, and R.-G. Avid, "Electromyography Signal Acquisition and Analysis System for Finger Movement Classification," 2019.

[3] K. Hartman, "Getting Started with MyoWare Muscle Sensor," 2018.

[4] J. G. Abreu, J. M. Teixeira, L. S. Figueiredo, and V. Teichrieb, "Evaluating Sign Language Recognition Using the Myo Armband," Proc.-18th Symp. Virtual Augment. Reality, SVR 2016, pp. 64–70, 2016.

[5] E. Lopez-Larraz, O. Mozos, J. Antelis, and J. Minguez, "Syllable-based speech recognition using auditorylike features," 32nd Annu. Int. Conf. IEEE EMBS, 2010.

[6] S. Pancholi and A. M. Joshi, "Time Derivative Moments Based Feature Extraction Approach for Recognition of Upper Limb Motions Using EMG," IEEE Sensors Lett., vol. PP, no. 3, pp. 1–1, 2019.

[7] C. Savur and F. Sahin, "American Sign Language Recognition system by using surface EMG signal," 2016 IEEE Int. Conf. Syst. Man, Cybern. SMC 2016 - Conf. Proc., pp. 2872–2877, 2017.

[8] S. Pancholi and R. Agarwal, "Development of low cost EMG data acquisition system for arm activities recognition," 2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2016, pp. 2465–2469, 2016.

[9] R. Zhou, Q. Luo, X. Feng, and C. Li, "Design of a wireless multi-channel surface EMG signal acquisition system," 2017 3rd IEEE Int. Conf. Comput. Commun. ICCC 2017, vol. 2018-Janua, pp. 279–283, 2018.

[10] M. Tariquzzaman, F. Khanam, M. H. A. Sohag, and M. Ahmad, "Design and implementation of a low cost multichannel rectified EMG acquisition system," 19th Int. Conf. Comput. Inf. Technol. ICCIT 2016, pp. 261–265, 2017.

[11] H. Ghapanchizadeh, S. A. Ahmad, and A. J. Ishak, "Developing multichannel surface EMG acquisition system by using instrument opamp INA2141," pp. 258–263, 2014.

Implementation of a Beowulf Cluster and Analysis of its Performance in Applications with Parallel Programming

Enrique Lee Huamaní[1], Patricia Condori[2], Avid Roman-Gonzalez[3]
Image Processing Research Laboratory (INTI-Lab)
Universidad de Ciencias y Humanidades
Lima, Perú

***Abstract*—In the Image Processing Research Laboratory (INTI-Lab) of the Universidad de Ciencias y Humanidades, the permission to use the embedded systems laboratory was obtained. INTI-Lab researchers will use this laboratory to do different research related to the processing of large scale videos, climate predictions, climate change research, physical simulations, among others. This type of projects, demand a high complexity in their processes, carried out in ordinary computers that result in an unfavorable time for the researcher. For this reason, one opted for the implementation of a high-performance cluster architecture that is a set of computers interconnected to a local network. This set of computers tries to give a unique behavior to solve complex problems using parallel computing techniques. The intention is to reduce the time directly proportional to the number of machines, giving a similarity of having a low-cost supercomputer. Different performance tests were performed scaling from 1 to 28 computers to measure time reduction. The results will show if it is feasible to use the architecture in future projects that demand processes of high scientific complexity.**

Keywords—High-performance cluster; distributed programming; computational parallelism; Beowulf cluster; high-efficiency computing

I. Introduction

With the growth of technological advances related to the world of computing, new techniques are emerging that take full advantage of computers that are interconnected by the same local network. The idea is to meet specific needs in less time than an ordinary computer. Performing processes of high availability, efficiency, and performance has become critical to providing better services, optimizing time resulting from complex problems and having continuous availability.

In this work is carried out the implementation of a high-performance cluster type Beowulf, which is the set of low-cost computers interconnected by a network to solve problems of high scientific complexity in less time [1]. This work was done due to the proposal of new projects where the time of the result takes hours or days depending on the complexity of the algorithm. Thanks to this work, one is able to reduce the time in providing results using distributed programming.

The work was done in the embedded systems laboratory of the University of Sciences and Humanities for scientific purposes that would benefit the INTI-Lab to perform simulations of high scientific complexity. These types of architectures are commercially considered supercomputers because they are designed to increase computing power by allowing parallel processing of tasks and high-speed communication [2]. Twenty-eight computers will be used, of which a performance test will be performed using a specialized algorithm. The idea is to measure their scalability and determine the exact number of computers to be used due to bottlenecks that can occur in the communications network.

The computers used in the architecture will continue being of first use for the university student. For that reason, there is a calendar for its scientific use without causing inconveniences at the time of using the laboratory. Reusing the hardware resources for the implementation of this work is not a new idea. One has the Polytechnic University of Altamira [3] that uses the five computers of the optimization laboratory and networks to solve complex problems with a specialized package to measure its scalability. Recycled computers should not be wasted, and a focus can be given to their use in a cluster architecture. As is the case of the National Engineering University [4], which uses recycled computers to obtain a maximum computational benefit. All these architectures that were mentioned are called Beowulf clusters. They are so-called because of the low or regular computing resource [5]. They will use four elements of hardware that are the RAM, the central processing unit (CPU), and its network card.

In this work, the maximum potential of the computers of the laboratory of embedded systems will be used, being the completion of this work a supercomputer with low computer resources.

II. Methodology

The cluster architecture will be implemented in Laboratory 302-B that will use the 29 computers interconnected to a switch. A computer will be in charge of distributing the problem to the 28 remaining computers that will solve a problem applying techniques of computational parallelism. The characteristics of hardware are of low cost; for that reason, the name of cluster Beowulf is chosen that will be detailed with precision in another section of this work. As can be seen in Fig. 1, in the lower-left is specified in the operating system, and the tool one is using for the realization of parallelism between the nodes because they are computers also used for the academic field all will be connected directly to the public network. Concerning the issue of security from the orchestrator computer will generate a unique security key thanks to the Secure Shell protocol, which will be copied to each laboratory computer to prevent an external agent cannot access not having access permissions.

A. Cluster Arquitecture

The cluster architecture can be of three types: high performance cluster (HPC), high availability cluster (HA) and high efficiency cluster (HP) [6] in this work an HPC architecture was implemented that uses powerful tools and processes of computing to generate data in advanced academic research [7], this type of architecture was chosen to take maximum advantage of computing resources in order to obtain successful results in less time, the more project proposals there are the need to use other types of architecture making the future laboratory of the institution a hybrid cluster, the characteristics of the HPC of the laboratory are as follows.

1) Master node: It is the computer in charge of distributing the problem applying parallel programming to distribute it to the slave nodes in order to give a result in less time, in it, one can see the ecosystem of the architecture installing some monitoring packages.

2) Slave node: These computers have two functions, one of them is to take a portion of the general problem that is distributed by the master node and return the final result when it finishes processing, they do not need to have a graphical user interface because it basically needs to be connected to the network to extract the number of cores for the process.

3) Communication network: It is the means that will help the communication between the slave nodes and the master. The better the communication by the network equipment, the lower the network traffic, making the performance of the processes more favorable.

4) Secure shell protocol: In this work, one uses the Secure Shell (SSH) thanks to this protocol the master node can interact securely with the slave node, for information security reasons a unique key was generated from the master node and copies were made to the slave nodes to have a secure cluster architecture.

5) Paralleling tools: To make computers work in parallel requires specialized tools, in this case, was used Open Mpi which is an implementation of open-source message step interface[1], this tool will help the realization of computational parallelism techniques, in the taxonomy of Flynn extracted from [8] shows two types of parallelism.

a) MISD: Applies the technique of multiple instructions to data where its functional units perform different operations on the same data.

b) MIMD: Applies multiple instruction techniques, multiple data used to achieve parallelism, the machines that use these techniques have several processors that operate asynchronously and independently.

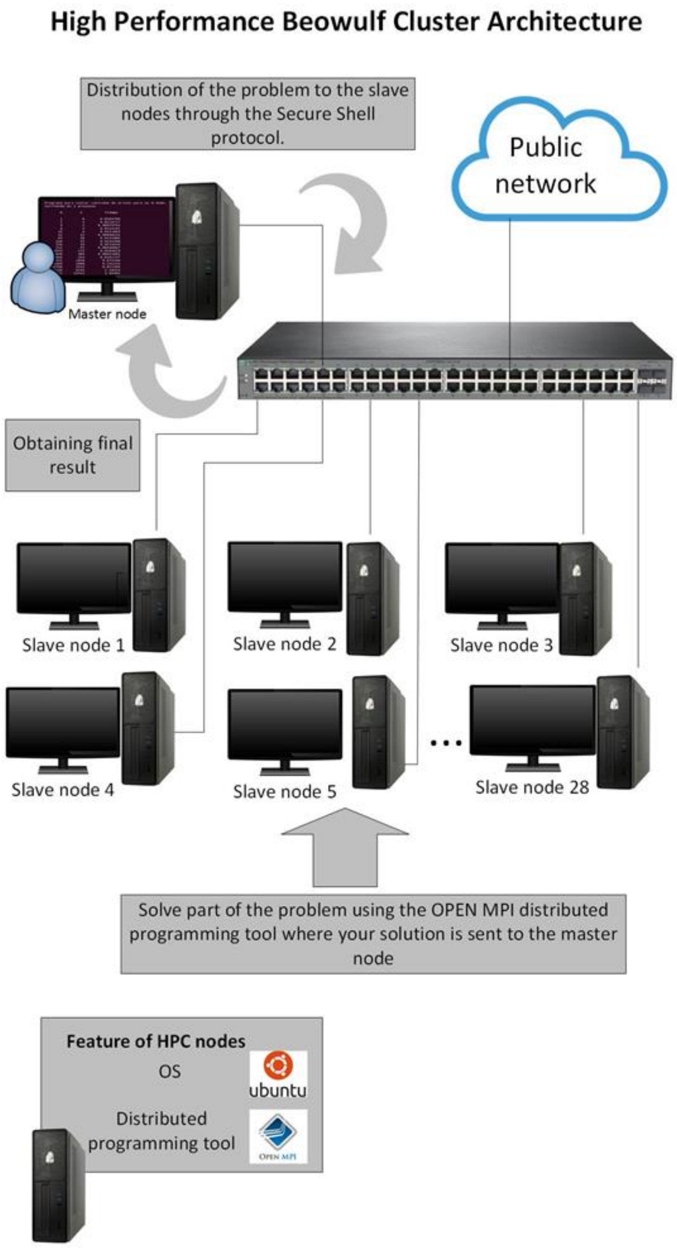

High Performance Beowulf Cluster Architecture
Distribution of the problem to the slave nodes through the Secure Shell protocol.
Public network
Master node
Obtaining final result
Slave node 1
Slave node 2
Slave node 3
Slave node 4
Slave node 5
...
Slave node 28
Solve part of the problem using the OPEN MPI distributed programming tool where your solution is sent to the master node
Feature of HPC nodes
OS
ubuntu
Distributed programming tool
Open MPI

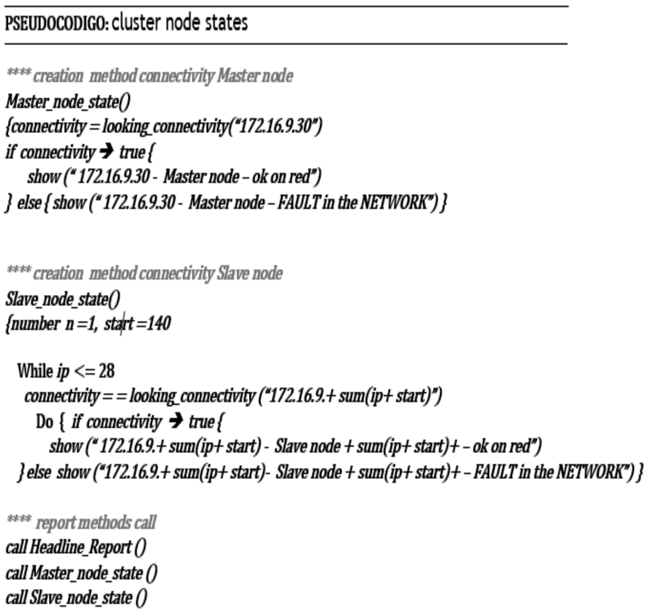
Fig. 1. Design of the Beowulf Cluster Architecture.

B. Architectural Status Monitoring

One of the most common cases in the use of a Beowulf cluster is the disconnection that can occur in some of the computers, to access to each one of it to make sure that it has communication can be a tedious work even more in the case of 29 computers, for this problem an algorithm was developed in the programming language Python that shows a report of the nodes with their state of connectivity in Fig. 2 the pseudo code is shown.

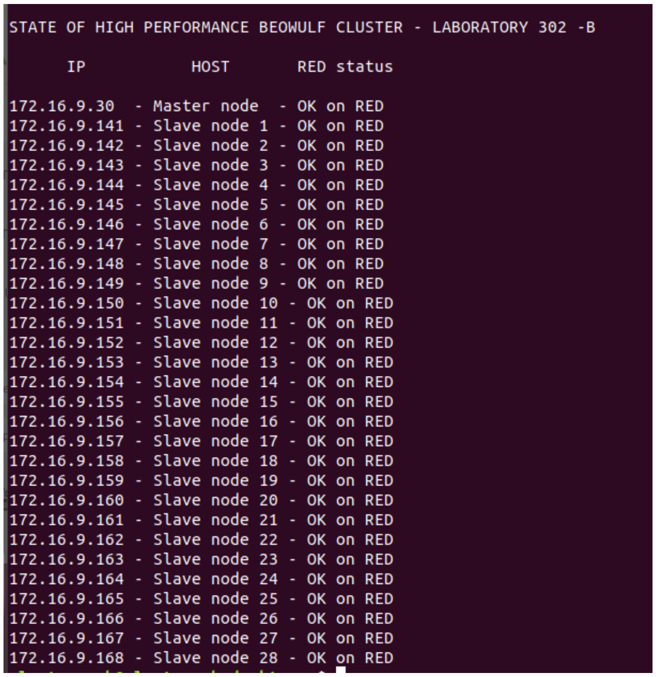
As a result, we have the following report, as shown in Fig. 3 that shows 28 slave nodes and one master node, all activated in a network.

[1] The Open MPI Project, “A High Performance Message Passing Library” 2019. [Online]. Available: https://www.open-mpi.org/



```
PSEUDOCODIGO: cluster node states

**** creation  method connectivity Master node
Master_node_state()
{connectivity = looking_connectivity("172.16.9.30")
if connectivity ➔ true {
    show (" 172.16.9.30 -  Master node – ok on red")
} else { show (" 172.16.9.30 -  Master node – FAULT in the NETWORK") }


**** creation  method connectivity Slave node
Slave_node_state()
{number  n =1,  start =140

  While ip <= 28
   connectivity = = looking_connectivity ("172.16.9.+ sum(ip+ start)")
     Do {  if connectivity ➔ true {
        show (" 172.16.9.+ sum(ip+ start) -  Slave node + sum(ip+ start)+ – ok on red")
  } else  show ("172.16.9.+ sum(ip+ start)-  Slave node + sum(ip+ start)+ – FAULT in the NETWORK") }

**** report methods call
call Headline_Report ()
call Master_node_state ()
call Slave_node_state ()
```

Fig. 2. Pseudo Code Node State of the Beowulf Cluster.



```
STATE OF HIGH PERFORMANCE BEOWULF CLUSTER - LABORATORY 302 -B

      IP            HOST       RED status

172.16.9.30  - Master node  - OK on RED
172.16.9.141 - Slave node 1 - OK on RED
172.16.9.142 - Slave node 2 - OK on RED
172.16.9.143 - Slave node 3 - OK on RED
172.16.9.144 - Slave node 4 - OK on RED
172.16.9.145 - Slave node 5 - OK on RED
172.16.9.146 - Slave node 6 - OK on RED
172.16.9.147 - Slave node 7 - OK on RED
172.16.9.148 - Slave node 8 - OK on RED
172.16.9.149 - Slave node 9 - OK on RED
172.16.9.150 - Slave node 10 - OK on RED
172.16.9.151 - Slave node 11 - OK on RED
172.16.9.152 - Slave node 12 - OK on RED
172.16.9.153 - Slave node 13 - OK on RED
172.16.9.154 - Slave node 14 - OK on RED
172.16.9.155 - Slave node 15 - OK on RED
172.16.9.156 - Slave node 16 - OK on RED
172.16.9.157 - Slave node 17 - OK on RED
172.16.9.158 - Slave node 18 - OK on RED
172.16.9.159 - Slave node 19 - OK on RED
172.16.9.160 - Slave node 20 - OK on RED
172.16.9.161 - Slave node 21 - OK on RED
172.16.9.162 - Slave node 22 - OK on RED
172.16.9.163 - Slave node 23 - OK on RED
172.16.9.164 - Slave node 24 - OK on RED
172.16.9.165 - Slave node 25 - OK on RED
172.16.9.166 - Slave node 26 - OK on RED
172.16.9.167 - Slave node 27 - OK on RED
172.16.9.168 - Slave node 28 - OK on RED
```

Fig. 3. Beowulf Cluster Status Result.

C. *Cluster Beowulf*

The implementation of the work is of Beowulf type that is denominated with this prefix for the use of components of hardware of low cost that behave as if they were an only computer [9] the computers of the laboratory of embedded systems are used, space where the students of the university carry out their academic activities, therefore, the laboratory has a particular schedule for the accomplishment of investigations, in this architecture was used a number of 28 slave computers and a master computer using a number of 196 cores that will process the problem applying parallel techniques computational parallelism distributing the problem to each of its cores for obtaining results in less time, Fig. 4 shows the computers used in the performance testing process.

All equipment has the same hardware characteristics, as shown in Table I.

D. *Performance Testing*

In this work, performance tests were performed to measure scalability with an intensive calculation algorithm for the sum of prime numbers used in the C++ programming language.

1) Parallel algorithm for calculating prime numbers: For the performance tests, an algorithm used in previous work with virtual machines was selected [10] that applies parallelism techniques thanks to Open Mpi that performs intensive iterations to each of the numbers to validate if it is a prime number. In this work, an intensive calculation is carried out by testing 2, 4, and 8 million iterations, Fig. 5 shows the pseudocode.

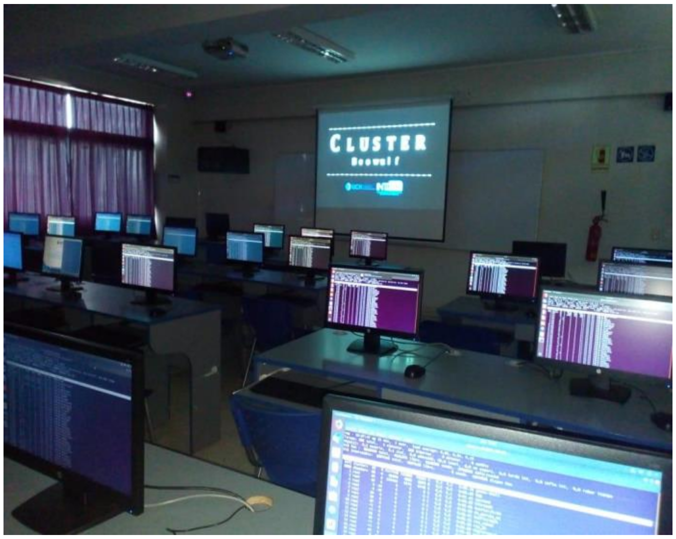
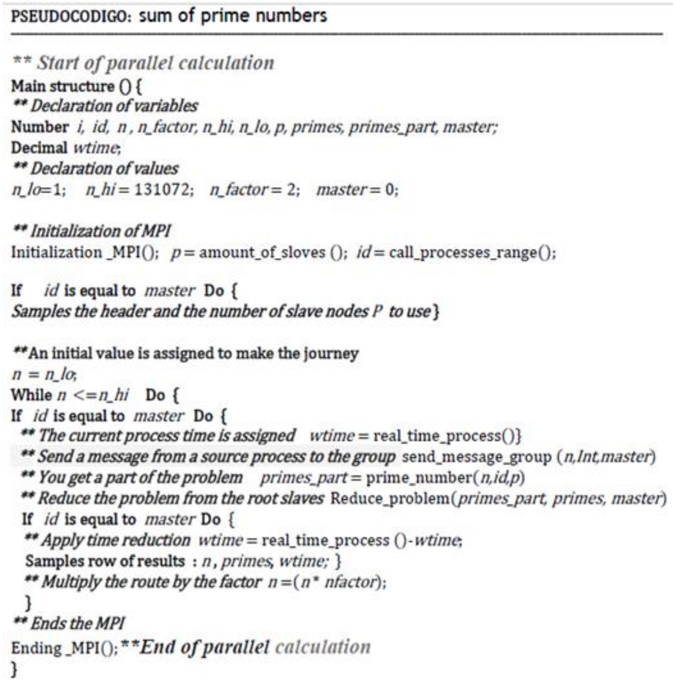


Fig. 4. Comparison of the Performance Test.

TABLE. I. HARDWARE CHARACTERISTICS OF THE BEOWULF CLUSTER COMPUTERS

	Description
Modell	HP EliteDesk 800 G1 SFF
HDD	1 TB
RAM	8 GB
Processor	Intel® CoreTM i7-4790 CPU
Total Cores	7
Type of Operating System	64-bit
Operative System	Ubuntu 18.04



```
PSEUDOCODIGO: sum of prime numbers

** Start of parallel calculation
Main structure () {
** Declaration of variables
Number i, id, n , n_factor, n_hi, n_lo, p, primes, primes_part, master;
Decimal wtime;
** Declaration of values
n_lo=1;   n_hi = 131072;   n_factor = 2;   master = 0;

** Initialization of MPI
Initialization _MPI();   p = amount_of_sloves ();  id = call_processes_range();

If    id is equal to  master  Do {
Samples the header and the number of slave nodes P  to use}

**An initial value is assigned to make the journey
n = n_lo;
While n <=n_hi   Do {
If  id is equal to  master  Do {
  ** The current process time is assigned   wtime = real_time_process()}
  ** Send a message from a source process to the group send_message_group (n,Int,master)
  ** You get a part of the problem    primes_part = prime_number(n,id,p)
  ** Reduce the problem from the root slaves Reduce_problem(primes_part, primes, master)
  If  id is equal to  master Do {
   ** Apply time reduction  wtime = real_time_process ()-wtime;
   Samples row of results  : n , primes, wtime; }
   ** Multiply the route by the factor  n =(n * nfactor);
   }
** Ends the MPI
Ending _MPI();**End of parallel calculation
}
```

Fig. 5. Pseudo Code of the Cousin Calculation Algorithm.

III. Result

In this section, performance tests are shown using the algorithm of the calculation of prime numbers using distributed programming. Scalability measurements are made using from one to 28 slave nodes of the Beowulf cluster.

For the execution of the algorithm, we will use a line of code from the console of the master node that is: mpirun –np # -hostfile ../.mpi_hostfile ./primos, where the symbol # represents the number of cores and .mpi_host file the number of slave nodes and ./primos the execution of the compiled algorithm. The more slave nodes used, the more kernels must be included from the Linux console to deliver the results in less time.

In Table II, the first tests are performed with 2 million iterations using from one to 28 slave nodes. In Fig. 6, the same result is shown in the form of a statistical graph, taking into consideration the number of slave nodes versus the time of the result.

In Table III, the tests are performed with 4 million iterations using from one to 28 slave nodes. In Fig. 7, the same result is shown in the form of a statistical graph, taking into consideration the number of slave nodes versus the time of the result.

In Table IV, the tests are performed with 8 million iterations using from one to 28 slave nodes. In Fig. 8, the same result is shown in the form of a statistical graph, taking into consideration the number of slave nodes versus the time of the result.

Finally, Fig. 9 shows a statistical graph of the number of slave nodes versus the result time, taking as reference the three performance tests used in the previous tables.

TABLE. II. Intensive Calculation with 2 Million Iterations

Number of slave nodes	Time of result	Core
Slave node 1	92.0369	7
Slave node 2	97.8105	14
Slave node 3	46.722	21
Slave node 4	46.1112	28
Slave node 5	25.8577	35
Slave node 6	46.1088	42
Slave node 7	13.8339	49
Slave node 8	23.1545	56
Slave node 9	15.4168	63
Slave node 10	23.2592	70
Slave node 11	9.33184	77
Slave node 12	23.1032	84
Slave node 13	8.07527	91
Slave node 14	13.2397	98
Slave node 15	11.6152	105
Slave node 16	12.1778	112
Slave node 17	6.44756	119
Slave node 18	15.3812	126
Slave node 19	5.85358	133
Slave node 20	11.5912	140
Slave node 21	10.04327	147
Slave node 22	11.2397	154
Slave node 23	8.14788	161
Slave node 24	4.75896	168
Slave node 25	6.44756	175
Slave node 26	4.48963	182
Slave node 27	3.85358	189
Slave node 28	2.5912	196

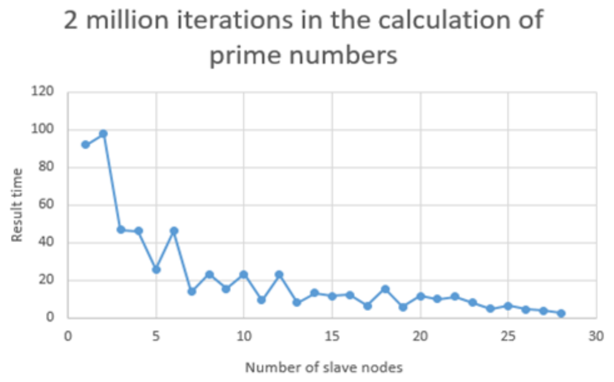


Fig. 6. Statistical Diagram of the Calculation of 2 Million Iterations.

TABLE. III. INTENSIVE CALCULATION WITH 4 MILLION ITERATIONS

Number of slave nodes	Time of result	Core
Slave node 1	351.2	7
Slave node 2	351.09	14
Slave node 3	180.775	21
Slave node 4	178.562	28
Slave node 5	91.6705	35
Slave node 6	176.087	42
Slave node 7	53.1501	49
Slave node 8	88.187	56
Slave node 9	58.6974	63
Slave node 10	88.4901	70
Slave node 11	35.5622	77
Slave node 12	88.2121	84
Slave node 13	31.1266	91
Slave node 14	50.4507	98
Slave node 15	44.3467	105
Slave node 16	44.1493	112
Slave node 17	24.7582	119
Slave node 18	61.3337	126
Slave node 19	22.0287	133
Slave node 20	44.1297	140
Slave node 21	31.1266	147
Slave node 22	22.4177	154
Slave node 23	25.3467	161
Slave node 24	27.4675	168
Slave node 25	19.7582	175
Slave node 26	15.3337	182
Slave node 27	17.6287	189
Slave node 28	13.56297	196

TABLE. IV. INTENSIVE CALCULATION WITH 8 MILLION ITERATIONS

Number of slave nodes	Time of result	Core
Slave node 1	351.2	7
Slave node 2	351.09	14
Slave node 3	180.775	21
Slave node 4	178.562	28
Slave node 5	91.6705	35
Slave node 6	176.087	42
Slave node 7	53.1501	49
Slave node 8	88.187	56
Slave node 9	58.6974	63
Slave node 10	88.4901	70
Slave node 11	35.5622	77
Slave node 12	88.2121	84
Slave node 13	31.1266	91
Slave node 14	50.4507	98
Slave node 15	44.3467	105
Slave node 16	44.1493	112
Slave node 17	24.7582	119
Slave node 18	61.3337	126
Slave node 19	22.0287	133
Slave node 20	44.1297	140
Slave node 21	31.1266	147
Slave node 22	22.4177	154
Slave node 23	25.3467	161
Slave node 24	27.4675	168
Slave node 25	19.7582	175
Slave node 26	15.3337	182
Slave node 27	17.6287	189
Slave node 28	13.56297	196

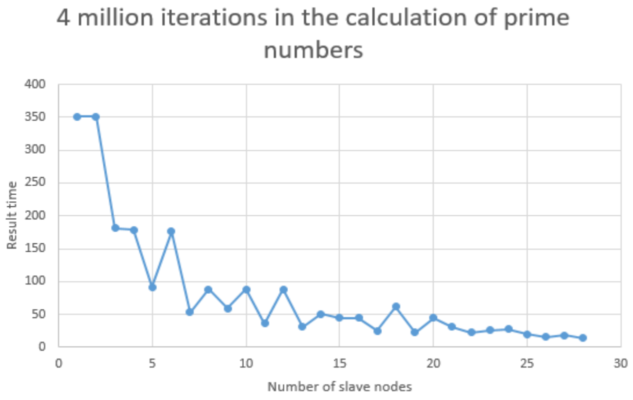


Fig. 7. Statistical Diagram of the Calculation of 4 Million Iterations.

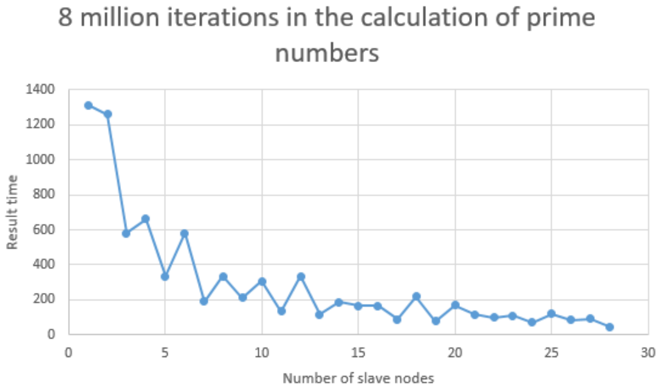


Fig. 8. Statistical Diagram of the Calculation of 8 Million Iterations.

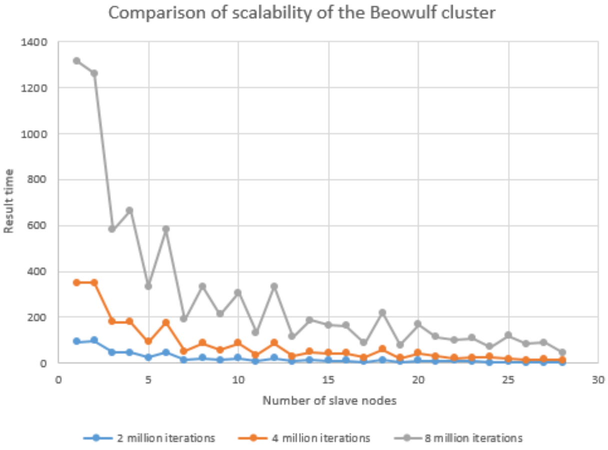


Fig. 9. Comparison of the Performance Test.

IV. Discussion and Conclusions

The work serves as a starting point for the realization of algorithms of high scientific complexity. It has scheduled a schedule of continuous improvement where the activities will be carried out depending on the need that arises in the direction of research of the Universidad de Ciencias y Humanidades. Improvements will include high availability to obtain large volumes of information using Big Data techniques as it does in [11]. This work has similarity concerning the measurement of scalability to use more nodes that demonstrate the efficiencies of these HPC architectures with Big Data Hadoop. Concerning the results section, one sees a reduction in time when more odd nodes are used. This situation is due to its cores that carry out the work in parallel. This work can be improved using a higher number of cores without having to resort to using a new slave node, as one has the case of [12]. This work of the Universidad Nacional de Ingeniería that takes full advantage of the GPU that each computer has demonstrated that the performance is five times higher compared to using CPU.

In future works, related to the increase of the Beowulf cluster potential, it will be proposed to include graphics cards. These graphic cards will make this architecture more powerful using PyCuda.

This work demonstrates that the use of a Beowulf cluster architecture using embedded systems laboratory computers reduces time without the need to acquire specialized equipment. As shown in the results section of Fig. 9, the more complex the problem, the more efficient the slave nodes will be, concluding that the implementation of this architecture meets the proposed objectives.

References

[1] Jiménez and A. Medina, “Cluster de Alto Rendimiento,” Fac. Ing.-UMSA Clust., vol. 1, no. 1, 2014.

[2] J. A. Fiestas-Iquira, “El papel de la supercomputación en la investigación: astrofísica de núcleos galácticos y agujeros negros,” Interfases, vol. 0, no. 008, p. 49, 2015.

[3] M. Vargas-Martínez, S. Gómez-carpizo, J. Sandoval-Sánchez, and G. Castillo-Valdez, “Revista de Sistemas Computacionales y TIC’ s Construcción de clusters de computadoras de bajo costo utilizando software libre Revista de Sistemas Computacionales y TIC ’ s,” Rev. Sist. Comput. y TIC’s, vol. 2, no. 4, pp. 19–25, 2016.

[4] J. Fiestas and C. M. Cruz, “Construcción e Implementación de un Clúster con máquinas PCs recicladas.,” Rev. la Fac. Ciencias la UNI, vol. 14, no. 1, 2014.

[5] R. Samir and R. Caro, “Implementación De Un Clúster Experimental Bajo,” p. 12, 2014.

[6] D. Armando et al., “Computación De Alto Desempeño Para Cálculos De Química Mecano-Cuántica,” p. 3, 2015.

[7] L. Chuquiguanca, E. Malla, F. Ajila, and R. Guamán-quinché, “Arquitectura clúster de alto rendimiento utilizando herramientas de software libre,” vol. 2, no. 1, pp. 1–8, 2015.

[8] A. † Velarde-Martinez, Luna-Ramirez, E. & Haro-Hernandez, and José, “Liebres Inteligentes: Sistema de Multicomputadoras para el procesamiento paralelo de aplicaciones científicas,” Rev. Tecnol. e Innovación, vol. 2, no. 3, pp. 454–463, 2015.

[9] P. Burhanuddin, N. Universitas, M. Indonesia, P. R. View, and P. B. Nuhung, “Cluster Computing Analysis Based on Beowulf Architecture,” Int. J. Comput. Informatics, vol. 1, no. April, pp. 9–15, 2016.

[10] E. L. Huamaní, P. Condori, and A. Roman-gonzalez, “Virtualizing a Cluster to Optimize the Problems of High Scientific Complexity within an Organization,” vol. 10, no. 6, pp. 618–622, 2019.

[11] D. Schmidt, W. C. Chen, M. A. Matheson, and G. Ostrouchov, “Programming with BIG Data in R: Scaling Analytics from One to Thousands of Nodes,” Big Data Res., vol. 8, pp. 1–11, 2017.

[12] N. M. Lapa Romero, J. A. Fiestas Iquira, A. Tenorio Trigoso, and Y. Nuñez Medrano, “Pruebas de rendimiento sobre el Clúster de CPUs y GPUs empleando simulación N-body,” no. July, pp. 19–21, 2018.

An Internet of Things (IOT) based Smart Parking Routing System for Smart Cities

Elgarej Mouhcine[1], El Fazazi Hanaa[3],
Khalifa Mansouri[4], Youssfi Mohamed[5]
Laboratory SSDIA
ENSET University Hassan II
Mohammedia, Morocco

Karouani Yassine[2]
Laboratory RITM
ENSEM-ESTC-UH2C
Casablanca, Morocco

Abstract—**Recently, the number of cars on the road has been growing due to the increase in car manufacturing in parallel with customer services provided to help the new driver to buy cars at affordable prices. On the other hand, we find that the infrastructure of big cities cannot support this number of cars and with the disorganization of parking places in the city this problem will lead us to have serious problem in the city which involves the increase of drivers requests to find the nearest parking places to avoid traffic congestion in those areas. In parallel today we talk about the concept of smart cities and how can we use the evolution of the Internet of Things (IoT) to improve the quality of the smart city. Several efforts have been made on the Internet of Things to improve the reliability and the productivity of public infrastructure. Many problems have been handled and controlled by the IoT such as vehicle traffic congestion, road safety and the inefficient use of car parking spaces. This work introduces a novel technique based on a distributed cloud architecture of IoT to manage the parking systems combined with a distributed swarm intelligence technique using the Ant System algorithm to improve the process of finding the nearest car parking in the minimum time based on the state of traffic on this road. This prototype will help drivers to find the nearest car parking and improve the exploitation of the available car parking in the city.**

Keywords*—*Intelligent traffic system; internet of things; swarm intelligent; ant colony optimization; vehicle routing system; multi-agent system; smart parking system; smart cities; cloud computing system

I. Introduction

The theory of internet of things (IoT) began with objects as a pre-identified devices. Those devices can be controlled, tracked and monitored by using a remote system connected through the internet. IoT is based on the internet that will be used as a link of communication between things and physical objects (devices) through a network of communication. This new concept is based on two words "internet" and "things", the first element presents the most tool of communication between several entities (computers, phones, tablets, servers, etc.) based on a set of protocols of communications to achieve the processes of (in/out) of data between several entities. The second word "Things" presents a physical or logical devices or objects viewed as a final device on this network system. IoT, in general, presents a set of devices used for data transfer between a set of entities toward organizing and managing units to analyze the received data. These things are combined with embedded devices that allow them to compute and communicate with the available devices in the network. Actually, a city can be viewed as a smart city [8], [19], [20] by making the majority of his areas smart, such as water supply distribution, electricity distribution policy, smart transport distribution [11], [23] and smart health-care vehicles. Making the majority of these entities smart will lead us to build a smart city. Vehicle traffic control [1], [14] is one of the famous problems faced by each city. Hence, several works have been done to control and manage traffic jam based on traffic signal control [2], smart parking distribution system [4], [17] and smart route navigation [6]. The idea of building a smart city is now possible with the evolution of the internet of things and one of the main issues of the smart city is to find a new concept to control traffic road [2], [22] and managing car parking places [5]. Recently, in big cities, the object of finding available parking space [18] is become a serious problem and even is become harder with the increase in the number of vehicles on the road. In order to solve this problem, the new smart cities should take this problem into consideration which leads us to decrease the searching time for parking slots and decrease the traffic flow on the route. In the case of parking systems, the driver needs some smart solutions that inform them about the available car parking around them based on the available embedded devices. Many modern cities have decided to implement a set of sensors and devices technologies based on the concept of IoT to control and manage all the traffic data on the city. Actually, several systems have been made for car parking system that allow the driver to receive real-time information about the state of parking slots and the available parking on a specific area in the city only based on the web pages or using mobile applications. Each parking, use a set of sensors to control and manage the state of parking slots.

In this wrok, a new concept of smart parking system will be developed to reduce the time consumed by the driver to find an available parking space to park his vehicle also to reduce the traffic flow in the city [18]. Based on such a system, the user will be able to find the nearest parking spaces and they will propose for them the shortest path (in terms of distance and road traffic) to arrive at the location of the parking. In this work, a distributed solution will be introduced to find the available parking slots for a driver based on the technology of IoT and propose to him the shortest path in terms of the distance and the traffic flow to arrive in the best conditions based on the technique of the ant colony optimization [26]. The system allows users to choose the optimal route to reach their destination by looking for the nearest car parking around of them.

The rest of this paper is organized as follows: Section II looks at previous works which are related to the parking routing system, Section III describes the IOT concept and background. Section IV formalizes the problem of parking routing system based on the services provided by the IOT technologies. The ACO concepts are described in Section V. The distributed solution for the parking system based on the technologies of IOT and the AS optimization method is shown in Section VI. Section VII looks at on our simulation results based on a multi-agent environment for the parking system (PS), the paper ends with a conclusion and offers some research direction for the future.

II. Related Work

In this work [5], the authors propose a new advanced driver assistance systems (ADAs) which aim to detect vacant parking slots, then the selected parking will be recommended to this driver. The proposed system will start searching for the available parking slots and help the driver to arrive at this parking based on a guiding monitoring system dedicated to this task. When the driver arrives at the parking, the system will inform the others parking with the state of the newly available slots on this parking.

Other solutions have been made to solve the traffic jam problems by using the internet of vehicles combined with the cloud computing and the big data to control and manage the traffic routing system and made some optimal paths for vehicles and guiding them toward the available parking. In this work [24] a novel approach for routing vehicles to reach their destinations including parking spaces in the minimum time. The proposed system provides an intelligent travel guiding system that helps the driver in the urban area and avoid routes with a higher level of traffic congestion.

A genetic algorithm (GA) has been implemented in the parking routing system to provide a new strategy based on the behaviour of the GA to solve the problem of the parking systems. Authors in paper [25], create a new intelligent solution to find a series of best spaces of the moving cars at a parking space. With this method, they can route the vehicle inside the parking to achieve the best free slot in the minimum time.

The object of this paper [2] is to propose a new model that implement traffic system things to manage and control the state of the road network remotely. They propose a smart adaptive traffic system connected to a set of things using the internet which allows them to retrieve real-time information at each moment. The data collected from different components are controlled and analyzed by the traffic control office. The obtained data will allow as to evaluate the state of the road network at a specific area. Also, this system is able to help drivers by giving them real-time information about the traffic flow around this area.

In this work [3], authors try to give us the basic route to take into consideration when we implement the services of IoT, because users can access to anything only by using the Internet services over lightweight devices, also know all services are shared between user based on the new concept of cloud computing in which knowledge is moved away from hard devices like (Personal computer or mobile phones) to enormous computers named as datacenters. This architecture will help us to improve the quality of our services by allowing all users to access to data using the internet and the newly provided services are no more centralized because in CC we talk about distributing data between users using internet services.

III. IOT Literature and Background Review

The technology of IoT is playing an important role in the network of smart technologies, so it should be able to ensure a continuous connection between unlimited smart things to the internet. Based on such technology, we need to define a standard template for the IoT architecture that will be used to prepare a smart environment based on the technologies of IoT. In the architecture (Fig. 1), we have three levels defined on this system: Presentation, Middleware, Hardware level. Different architectures have been proposed in this context but all of them are based on these three layers. The Hardware level control and manage the set of physical sensors and embedded objects, the second level propose the set of components that will be used for the cloud computing and finally, the presentation level, used for proposing a visible vision for the final user based on some human interfaces, etc.

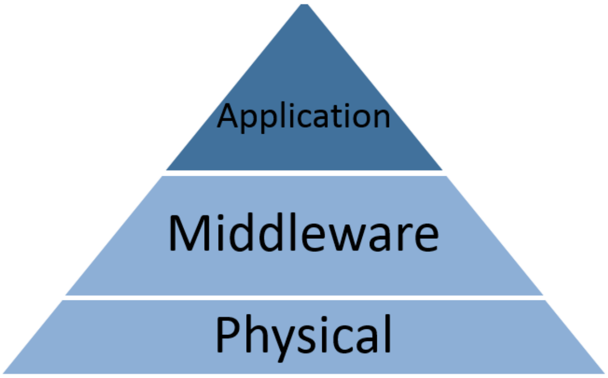


Fig. 1. The standard IoT architecture

A. Physical Level

The physical layer is based on a set of physical sensors and objects and the data received from those entities are collected to be analyzed and used for managing those entities. All the received information will be forwarded toward the middleware level using a secure channel of communication. Finally, all the data collected from those objects are stored and collected at this physical level.

B. Middleware Level

In this level, we will forward the data collected from the higher layer to the cloud system, we can use several technologies to transfer information such as 4G GSM, Wi-Fi, ZigBee, Ethernet, Bluetooth, etc. All the process is secured using TCP (Transmission Control Protocol) protocols and the data are transferred using HTTP (Hypertext Transfer Protocol) or FTP (File Transfer Protocol) mechanism. On the other hand, this layer has the ability to retrieve data from the cloud and help the physical layer in step of collecting the data. Based on this level, we can develop a set of applications which implement and re-use the received data to propose several solutions only based on the interconnected objects inside this platform.

C. Application Level

This level proposes a set of services for customers, it handles all the requests and the responses according to the parameters set for each request. We know that our architecture is based on cloud IoT system, so according to this request, this layer can search and find the set of objects that can give us such kind of information. This one is the main interface between human and cloud IoT. Actually, we see the combination of IoT with cloud computing to improve the quality of services provided by both of them. The IoT can improve his technologies such as the links of communication and storage process, energy consumption by implementing the unlimited resources of cloud computing. Basically, the platforms of cloud computing dedicated to the IoT technologies can be viewed as an intermediate actors between applications and things in the IoT network, in order to reduce all the functionalities and the complexities for preparing a good environment to run these applications. Several factors will push us to think about the strategy of combining the IoT with cloud computing such as:

- The Availability: with the new improvement on the cloud, the availability of resources are an important element especially with the cloud integration which aims to keep alive the set of resources running and up on time based on some services built for this job.
- Processing power: in the IoT environment we are using a set of sensors with limited processing capacity and when the sensor collect and capture the data it should be transferred to more powerful devices and objects where we can apply the rest of processing to retrieve the information needed from these things. With the help of cloud technologies, we can benefit from the real-time and the parallel processing performed by the cloud network.
- Middleware services: based on the cloud computing we can use the set of the link of communication used to transfer and tracking and managing objects from any where only using the internet (such as, 4G, Ethernet, Wi-Fi, etc.). With this remote technology, IoT can control and communicate with any types of things using remote communication.
- Storage limit: In the IoT architecture, we have a set of data collected from several things on the environment, we know that each thing has a limited storage capacity so the data collected or received by those devices should be transferred to the main agent that collect this information from various objects. then, the storage services provided by the cloud can help the IoT object to share and visualize the data from any places through specific web services.

IV. Overview of the System Architecture

The overview of the problem of parking routing system [18] can be described as showing in (Fig. 2), based on this architecture we can see that our system combines several entities which are distributed on the whole infrastructure and interconnected with each other based on the cloud computing platform. Some of these elements are viewed as intelligent components which are able to improve the quality of the collected data and send it to the next agent that will use this input data and propose a set of solutions for the final users. In the beginning, we need to identify all the set of entities that will work in our architecture such as parking places, road sensors, traffic road, road network, etc. Then we need to evaluate the state of the traffic road based on the set of pre-installed RSU (Road Sensor's Units) combined with the Google Traffic Services that allow us to retrieve the traffic flow on a specific area. In parallel with this process, we will find the optimal path [26] that will be followed by the driver to arrive at the nearest parking place to his location.

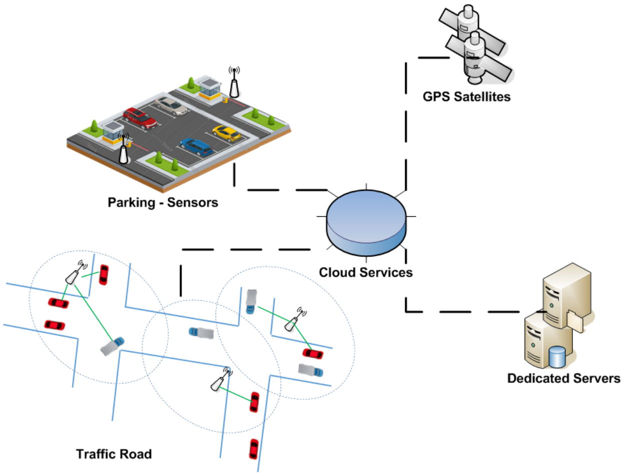


Fig. 2. Decentralized architecture for Managing the traffic road and parking system in a Smart City

A. Preparation of the Infrastructure

As shown in (Fig. 2), we need to locate all the available parking places and try to keep an eye on the state of available places at each parking, this task is done based on a set of sensors pre-installed on each parking which allow us to compute the number of available places on each parking, this collected data will be shared with a smart platform deployed on a distributed cloud system which allows us to centralize all these received data, we can identify the set of parking places on a specific area based on the classical clustering mechanism to select the set of nearest parking places to the position of the driver. In our case, each parking is controlled by an intelligent agent, this one is able to share real-update data about the state of slots on this parking at each moment. The cloud agent is able to communicate with these agents using TCP/IP network communication, based on this technology we can say that our infrastructure is remotely distributed.

B. Cloud Parking Routing System

When the cloud agent receives a request from the driver that search for a nearest parking slot, the system will catch the GPS coordinates of the driver and create an area around this position (Fig. 3), based on this area we can define the set of nearest parking places that are inside this area (we are based on the linear distance between the position of the driver and the pre-stored position of parking places in this area). When the set of available parking slots are defined, the system will prepare the path planning that will be followed by the driver toward one of those parking based on the state of the traffic road in this area and the total distance of the trip between the driver and the parking position. To get the state of traffic on this area, we are based on the RSU that allow us to retrieve the number of vehicles at a specific route and we call the Google Traffic services that give us a road network combined with traffic road. In this step, we can say that our road network for this specific area around the driver is completely prepared, then we will be able to call our distributed strategy to find the optimal path between the location of the driver and the set of parking places, in our problem the best path is viewed in terms of two elements: the total distance of the trip and the total time consumed to arrive at the location of the parking which is related to the traffic flow on this path. At the end of this process, our system will allow us to choose between a set of alternatives routes, each route will lead us to a specific parking in this area.

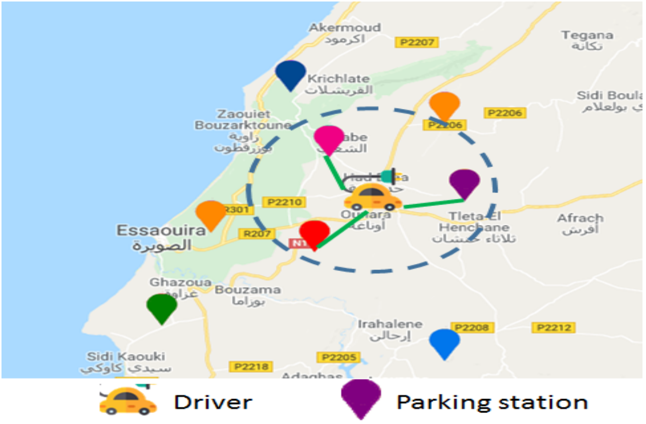


Fig. 3. Cloud Parking Clustering

V. The Principle of Basic Ant Colony Optimization

The ant colony optimization (ACO) technique is inspired by the behaviour of real ants in searching for food sources. Ants are initially started their journey by finding food sources randomly. If an ant finds a food source, it will leave a chemical substance named pheromone on the route used toward the source food when it returns back to their nest. It can appear that multiple ants find the same food source but with several trails. The amount of pheromone on routes will attract ants to use this trail. However, the best path or the shortest path will be more selected by the majority of ants and the amount of pheromone on this path will increase according to the number of ants that use this path. As time passes, we will see the majority of ants will take the same path to the food source which is considered as the optimal route to the given food.

The main objective of the ACO algorithms is to create a set of artificial ants inspired by the behaviour of real ants, they can build and find optimal solutions for combinatorial problems [9], [26], and it was used to solve complex problems [10] such as the travelling salesman problem (TSP), the vehicle routing problem (VRP). The artificial agents (ants) are created to find the best paths or optimal solutions. The process is based on sequential iterations to improve the quality of the proposed solution, also it takes into consideration the previous solutions to help agents to find optimal solutions in the next iteration. Many pieces of information are shared between ants by the proposed artificial pheromone, at each iteration a new amount of pheromone is updated according to the quality of each created solution. The algorithm monitoring the state of pheromone generated by each agent according to the goodness of the solutions. On the other hand, the amount of pheromone will decrease according to the evaporation process in order to avoid convergence to local optimum solutions. The system will repeat the same behaviour until the stop condition is met.

In the real world, ants are able to find the shortest path between their nest and food sources [21] by exploring the environment based on a chemical substance called pheromone, which is released by ants on the routes they travel. The proposed Ant System (AS) algorithm [15] is a procedure used to simulate the behaviour of real ants by producing a new mechanism for artificial ants that produces the same behaviour of real ants, with the addition of several attributes that helps the system to converge toward optimal solutions in minimum time. The different ACO algorithms depend on the nature of the problem and the solution for the problem [27], so the parameters of each algorithm are based on what the user needs to search for and how the nature of the environment will be used by applying the technique of the ACO. The AS is based on a set of predefined steps defined as follows (Algorithm 1):

Graph description: artificial entities will move between a set of discrete nodes in a discrete space. Since the problem solved by the AS algorithm is viewed as a discrete problem, they can be modelled by a graph built from nodes and edges.
Artificial ants preparation: a set of ants are distributed on the declared nodes in the proposed graph, the number of ants increases the collection of the generated solution since each ant is viewed as a candidate for the given problem.
Ants probability transition: ants are based on a probabilistic transition to move between vertex which is based on two main parameters, which includes the pheromone and a heuristic approach that defines the nature of the optimized problem.

The AS pseudocode paremeter desrciption:

- N_{nodes}: list of available nodes on the map.
- R_{routes}: list of available routes on the map.
- A_{ants}: number of ants used to find the bests solutions.
- $best_{route}$: contain the best solution computed after all ants finished their jobs.

Algorithm 1: AS pseudocode algorithme

```
Input: $N_{nodes}$,$R_{routes}$,$A_{ants}$
Result: $best_{route}$
foreach $A_i$ in $A_{ants}$ do
    foreach $N_i$ in $N_{nodes}$ do
        // Apply a probabilistic
           transition based on pheromone
           and heuristic parameter to
           select best next move and add
           this node to CR: collection
           of node
        $n$ = $select\_next\_node()$ ;
        $C_n = n$ ;
    end
    // Apply a local update of
       pheromone to the path visited
       by this ant
    $update\_pheromone(C_n)$ ;
    // Add constructed path to list of
       proposed routes
    $best_{route}$ = $C_n$ ;
end
```

- $select\_next\_node()$: this function implement the equation 1 to find the next node to visited (j) when the ant is on node i.
- C_n : represents the set of already visited nodes by the given ant.
- $update\_pheromone(C_n)$: apply the function 2 for updating the value of pheromone on the set of visited nodes and routes.

At the beginning of the search process, a consistent amount of pheromone is assigned to all arcs and the cost assigned to each arc can be defined according to the nature of the problem. For each ant in the colony, it will start its own trip by visiting the set of available nodes in this graph. Node selection will be based on a probabilistic transition by using the amount of pheromone and heuristic data according to the function $selec\_next\_node()$. The visited nodes will be appended to the trip followed by the ant. When the current ant visits all existing nodes in the graph, it will then apply a local update of pheromone$update_pheromone()$ by adding more pheromone on the set of routes visited during its search process. The current behavior will be repeated until all ants complete the existing nodes on the given graph. The system will allow us to select the best routes between the proposed solutions by ranking those solutions according to the length of each trip.

VI. The Parking Routing System Based on the IOT Combined with the Technique of the ACO

The main goal of our work is to help drivers to find the nearest parking places based on two parameters such as the distance between the current location of the vehicle and the parking slot combined with the state of the traffic on these proposed routes. To collect the state of the road we use a set of Road Sensors Units (RSU) that are able to compute the number of vehicles at a specific route and share this information with our cloud platform to be taken into consideration in the process of computing the optimal route for the driver. As shown in the previous section, the AS algorithm is based on the amount of pheromone to define the best routes visited by the majority of artificial ants. We will take the number of vehicles between two connected RSUs and correlate it with the amount of pheromone. So, if the number of vehicles is higher on this route so the amount of pheromone on this route will be the same. The number of vehicles passed between RSUs is automatically synchronized with the cloud system. This information will be used to evaluate the traffic flow between those intersections and we consider the number of stopping vehicles as the amount of pheromone on this route. This means that the pheromone density level is related to the number of non-moving vehicles on this route. The main objective of our work is to implement the behavior of the AS algorithm to search for all the best routes which have less number of vehicles (less traffic flow) and encourage the driver to go towards the route with less traffic flow. To better understand how we can implement the technique of the AS with the IOT technology [13], [16] we define the amount of the pheromone as the number of stopping cars in the line also we need to view the inverse of the original behavior of the AS algorithm by avoiding edges with higher amounts of pheromone and follow only routes with less number of vehicles (less amount of pheromone). Based on the state of the route received from the set of RSU and from the service of Google Traffic we can decide which route will be taken by the driver.

The road network will be viewed as a graph made from a set of intersections (nodes) and Routes (edges). In our case, nodes present the set of parking places and routes it will be the set of existing routes between those places. This graph will be presented as follows: $G = (I, R)$ Where I=(I_1,I_2,...,I_n) is the set of parking places and R is the set of routes. The main objective of this layout is to guide vehicles to the nearest parking in the fastest time while avoiding traffic on the road network. Avoiding traffic is done by skipping routes with a higher number of non-moving vehicles, and traffic flow delays. Each R_{ij} is identified by the distance d_{ij} and traffic flow T_{ij}. The time consumed to traverse R_{ij} is not dependent on the time consumed to traverse other routes because each route has its individual number of non-moving vehicles controlled by the RSU and the information provided by Google Traffic. The final goal is to find the optimal solution that minimizes two constraints, such as the distance and time consumed to finish the trip. So, in this problem, we will try to use two parameters (distance, traffic flow) to find the best path. The vehicle will apply a probabilistic transition to move from an intersection I_i towards I_j according to the rule below:

$$p_{ij}^{m}(it) = \frac{T_{ij}(it)^{\alpha} d_{ij}(it)^{\beta}}{\sum_{l \in \omega} T_{il}(it)^{\alpha} d_{il}(it)^{\beta}} \quad (1)$$

This is where the parameter (α, β) controls the influence of the traffic flow (pheromone) and the distance on the prepared solution. The parameter ω is the set of the non-visited intersections. The level of the traffic flow (pheromone) will evaporate (ρ rate of evaporation) depending on the information received from the RSU (the rate of the non-moving vehicles on this route). The traffic flow update is done according to this equation:

$$T_{ij} = (1-\rho).T_{ij} + \Delta_{ij} \tag{2}$$

The argument Δ_{ij} is related to the driver to see whether the vehicle has already visited this route R_{ij} . The new amount of traffic added on this route can be computed as follows:

$$\Delta_{ij} = \sum_{n=1}^{N} \frac{R_{ij}^{k}}{t_m(k)} \tag{3}$$

$tm_{(k)}$ represents the time needed to cross the distance d_{ij} of the route R_{ij} by the vehicle k based on the speed (ω_k) allowed on the route and the number of the available Non-moving cars.

$$tm_{(k)} = \frac{d_{ij}}{\omega(k)} \tag{4}$$

So, at each intersection, a vehicle will try to select the next intersection based on (Equation 1) and will then select this route based on two parameters: the distance and the traffic flow. The system will repeat the cycle until the destination is reached or the stop condition is met. At the end, the system will show the best routes that exists between the two locations. For each route, they will give us the total travel time consumed to cross this route with the total length (km) of this trip. (Equation 1) can be transformed into:

$$p_{ij}^{m}(it) = \frac{T_{ij}(it)^{\alpha} d_{ij}(it)^{1-\alpha}}{\sum_{l\in\omega} T_{il}(it)^{\alpha} d_{il}(it)^{1-\alpha}} \tag{5}$$

VII. A Distributed Strategy for the Parking Routing System

In this section, we describe the main architecture for the proposed distributed parking system which is based on several actors that collaborate together in asynchronous behaviour to build a smart parking system for vehicles in the city. The main actors that collaborate into our proposed system (Fig. 4) are a distributed agents each one of them works in a separate environment by sharing with other agents a set of parameters that are used to help drivers to find the best parking slot in the area.

A. Parking Sensors Units

In our parking infrastructure we are based on a set of sensors to control the number of vehicles (in/out) that visit our parking, also we use a set of infrared sensors under the category of Passive Infrared (PIF), those components are wirelessly connected to the Parking Agent (PA) which is modelled by a raspberry pi component using TCP/IP protocol which allows us to communicate using wifi network. So the PA will be able to communicate with those sensors and get the number of the available slots depending on the number of vehicles that leave the parking based on a simple equation by minus the number of vehicles inside the parking according to the number of vehicles that leave the parking from the outdoor sensor.

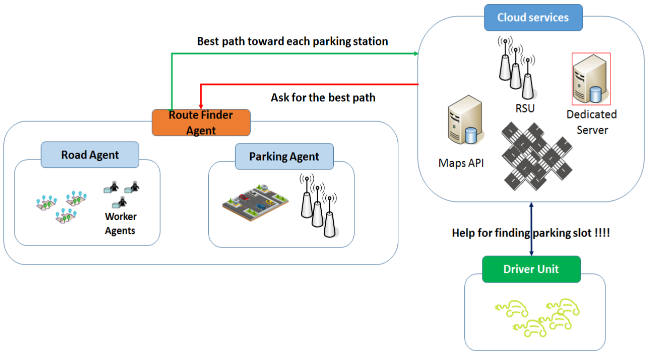


Fig. 4. Multi-agents architecture for managing the parking system based on the cloud control strategy

B. Parking Agent Unit

The Parking Agent Unit (PAU) is represented by a raspberry pi component, which is a microprocessor on a chip. This agent performs as an intermediate between the cloud routing system and sensors. All the sensors are remotely connected to the Path Finder Agent Unit. In this unit, we can find several GPIO (General Purpose Input/Output) pins, which act as an input data that allow us to connect all the available sensors in the parking. To control the process of communication between sensors and PAU we use an executable script writing with NodeJS programming language to read all the data received from the different GPIO and then transfer this data to the cloud routing system. So, we can say that our PAU will collect all data and share them with the cloud system. This data will be interpreted by other agents to select the best parking and compute the optimal path for vehicles to reach this parking.

C. Path Finder Agent and Cloud Services

In this cloud platform, we installed a dedicated server that acts as a receiver agent which collects all the data and saves it into NoSQL databases, in these databases we store all information about each parking agent such as the GPS coordinates of this parking and the real-time information about the available slots on this parking. In the other hand, we make a specific agent that creates a set of clustering area to group the nearest parking in a specific area, based on this clustering strategy the PFA can be able to fetch only the nearest area to the vehicle and optimizing the search process for the nearest parking. Also, the server can handle all the request of the users connected to our system that aims to find some available parking, according to each request, the system will invoke the PFA to respond on this demand and find an optimal solution to each request based on the available data on the server. In parallel, we define a cron job which is running at each new subscribed parking added to the list of available parking connected to our cloud system, this cron job execute a python script which create a set of clustering area each one of them contain the set of nearest parking, so at the end of this script we can see a set of groups, each one is viewed as an area that contains several parking. These areas will be used by the PFA to propose only the nearest parking to the driver and reduce the process of fetching for the nearest parking to the actual location of the driver.

As we can resume, the main function of this agent is to answer on two questions (Fig. 5) such as finding the set of parking that contains some available slots, to find the available parking we use the predefined clustering data performed by our cronjob, at the beginning the PFA will take the position of the vehicle and fetch in his local database for the nearest area to this location. Secondly, we help the driver to arrive at this parking in the minimum time and with a safety way by avoiding road with a huge traffic jam. To do this job, we are based on the technique of the AS to compute the optimal path that can be followed by the driver to arrive in the best conditions. So here we use a set of artificial ants named as worker agents (WA), each one of them will explore the graph (Fig. 6) made by our PFA, to build this graph we are based on the Google Maps API that helps us to create a map made from a set of intersection and several routes (edges) that are used to link between those intersections. To create this graph we are based on the actual location of the driver and the set of available parking places from the clustering done by our dedicated server, so based on this clustering we reduce the area that will be explored by our worker agents to find optimal routes between the driver and the set of nearest parking to the location of the vehicle. When the graph is made, each WA will try to visit the set of intersections between the location of the vehicle toward the proposed parking by the PFA, this work is done based on a probabilistic formula defined in the (Equation 1) which is based on two parameters are the linear distance and the level of traffic jam between the two adjacent intersections, when a WA visits all the available intersection it will be able to return back to the PFA with the optimal path that exists between the driver and the parking location. After all the worker agents complete their tours, the PFA will collect all the proposed paths and select the best one of them in terms of the distance and traffic jam, then this best route will be transferred to the driver to help them to arrive at this parking slot.

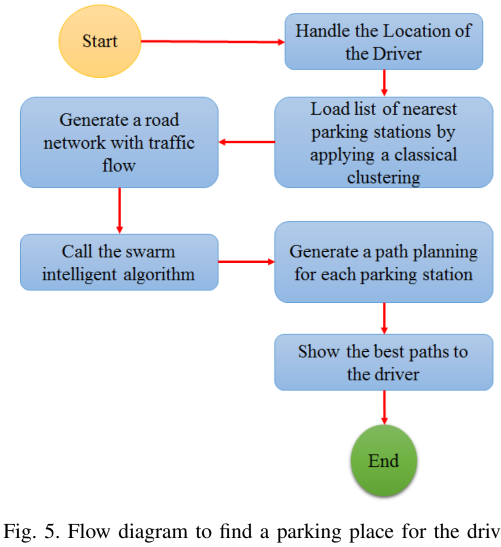

Start
Handle the Location of the Driver
Load list of nearest parking stations by applying a classical clustering
Generate a road network with traffic flow
Call the swarm intelligent algorithm
Generate a path planning for each parking station
Show the best paths to the driver
End

Fig. 5. Flow diagram to find a parking place for the driver

D. Vehicle Unit

This component is a hybrid application (Mobile/Web application) dedicated to being used as an interface of communication between the driver and the parking routing system through the cloud platform system. The application is made based on Ionic framework that combines Apache Cordova and AngularJS framework. This hybrid application can run on both mobile systems (Android/iOS). The application is connected to our back-office system deployed on the dedicated server, the process of communication is based on HTTP communication using JSON format to transfer data between these two entities. The driver can ask for finding the nearest parking place which contains some available slots and follows the best path proposed by the remote application to arrive on the minimum time and avoid traffic jams.

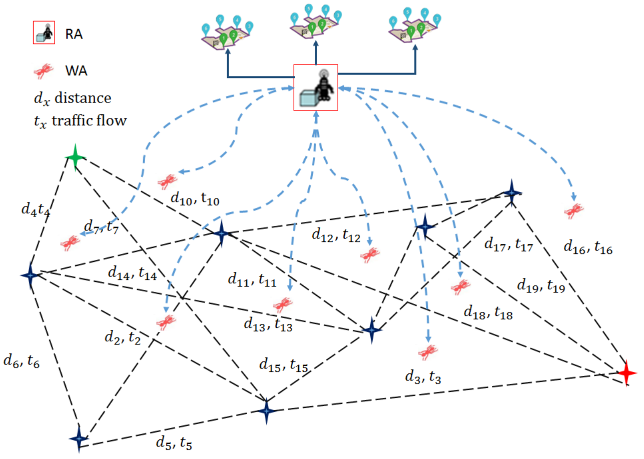


Fig. 6. Process of finding the best parking slot using the artificial ant agent

VIII. Simulation and Results

In this section, we present the main functions of our distributed solution which is divided into two components: front-end / back-end layer, the first one is used by the driver to fetch and locate the set of nearest parking stations, this side is represented by a web application developed based on the HTML 5 and AngularJS framework and to deploy this application we are based on Apache Tomcat container as a virtual server characterized by the following properties:

- Processor information: Intel(R) Xeon(R) CPU E5-1650 v2 @ 3.50GHz, 4 cores
- Kernel and CPU: Linux 4.13.8-1.el7.elrepo.x86_64 on x86_64
- Operating system: CentOS Linux 7.4.1708
- Apache Webserver: 2.4
- Java Version: 1.8
- Google Maps Javascript v3.1

The main package (Fig. 7) block of our proposed solution is the hardware package which contains (Raspberry PI, Arduino, IR sensors), in this package, our sensors are configured to sense their surroundings and notify the Arduino module with the real-time state of parking slots. The number of available parking slots are transmitted to the Arduino module. Our sensors are configured and displayed on each slot inside the parking building to keep an open eye on the state of the parking places.

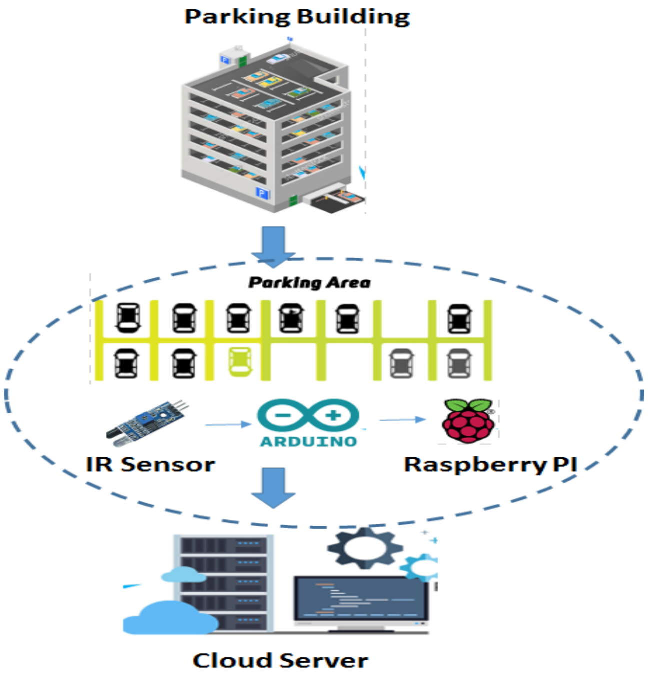


Fig. 7. Block diagram of the proposed smart parking system

To control the number of cars inside our parking building, we use another sensor placed at each entry of the building in the case when the parking has two entries. Each sensor will control the number of vehicles according to the triggering event when a vehicle passes through the detecting area of the sensor. To get the available parking slot, we placed at each slot an IR sensor, this one will be triggered when a car is inside his detecting area. After a car is detected by this sensor, it will send a notification push to the Arduino to update the state of this slot with two flags: occupied or available. On the other hand, we have the back-end side which is dedicated to compute and prepare all the data for the final user of this application, which is developed using the JADE (Java Agent Development Environment) that allow us to create a set of distributed agents, each one of them is able to work in parallel with other agents and has his private behaviour that presents the functions that will be done by this agent.

When the server handles the request of the driver, it will be able to create a circle around the GPS coordinates of the driver, then load all the existing parking around this circle, to fetch those parking stations we use our local database that contains the set of parking stations also we will show only parking with available slots, in our case, each parking station is viewed as a point-of-interest defined by the following attributes (latitude, longitude, available parking slots).

When the set of candidates parking stations are identified by our main agent, we can prepare our sub-graphs that contain the set of intersection between the actual location of the driver and the set of available parking stations. In Fig. 8, we can see the sub-road network created by the PFA, to put the state of traffic on this sub-area we are based on two remote services such as the RSU inside this area and the Google Traffic services. The data collected from the two services will be added to our road network, this data represents the cost of traffic at each route on these sub-graphs. At the end of this step, we can see that our graph is fully prepared by the set of intersections that can be visited by the drivers which are represented by the parking places and the actual location of the driver. All routes on this sub-road network are characterized by two attributes (the distance in kilometres and the number of non-moving vehicles). In the next step, the PFA will be able to run the distributed solution which allows us to create a set of artificial ants to find the best paths between the actual location of the driver and each selected parking station on this area. At the end, the web application will show for the driver the set of routes toward each parking and give him the total length of the trip and the approximate travelling time to reach the parking according to the state of traffic on this area. So here we allow the driver to choose one of those trips.

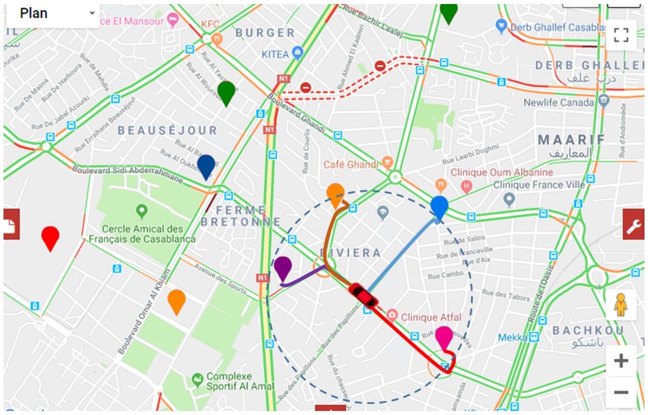


Fig. 8. Best path planning for a driver using the smart parking system strategy

Several scenarios have been made to see if our distributed strategy can be adopted in real life to help the driver into their daily needs. The first simulation has shown positive feedback and good results, but to see the efficiency of our method we should compare those results with other methods based on the same parameters to see if our strategy can give us useful feedback in terms of three constraints such as the length of the trip and the travelling time consumed on each trip and finally the time consumed by the prposed method to give as this solution (the best path between the driver location and the nearest parking station). As we can see that in this paper we are based on the techniques of swarm intelligent to build this solution, so we choose to compare our proposed solution with a combinatory method named as the Particle Swarm Optimization (PSO) [7] method to evaluate the results proposed by the two methods. The PSO was developed for continuous multi-objectives problems which try to find optimal solutions based on the collective behaviour of all those swarms methods [12], [28]. The PSO was proposed by Eberhart and Kennedy and was inspired by the behaviour of social bird flocking in nature. In the PSO algorithm, we are based on the mechanism of moving the better particles solutions locations in the solution space. The movement's system is controlled by the set of individuals and the best positions candidates. The PSO can propose the best solutions in a very fast convergence process due to a sample of two items such as memorizing individual best experiences (P_{best}) and information sharing of

the best global experiences (G_{best}). Also, the best-proposed solution G_{best} will be equal to the proposed solution of the best particle P_{best} in this solution space. Recently, this algorithm has been extended by adding the criteria of inertia weight.
The application of this algorithm has covered multiple domains especially the vehicle routing problem which aims to find the set of alternative routes for a given set of vehicles by generating a kind of path planning. In this algorithm we are based on a set of sequences iterations, each one of them has two steps such as clustering the set of heuristic input data in our case the set of parking stations by applying the district PSO algorithm and secondly we will be able to call the Simulated Annealing (SA) algorithm to help as in the process of discovering the space solutions. But, based on this mechanism we will spend much time in case of solving large scale of problems, so many improvements have been made on this algorithm to upgrade them to give as effective solutions for combinatorial problems. This improvement has been done by injecting another swarm intelligent methods such as ACO, Genetic Algorithm (GA), Tabu Search algorithm (TS), each one of those algorithms is used in a specific step on the PSO algorithm iterations.

TABLE I. Simulation parameters for the Distributed Parking System (DPS)

Parameter	Value
Development environment	Eclipse Luna & JADE
Number of request	$[10 - 200]$
Road Topology	$[5 - 60]$
Number of artificial agents	$[10, 30, 60]$
Rate of evaporation ρ	0.1
Parameter Q	1
Parameter β	0.1
Parameter α	0.2

To evaluate the performance of our proposed strategy, we are based on twelves scenarios (bins) (each scenario is viewed as a graph made from a set of parking stations and routes), for each one of those bins, we evaluate our proposed strategy compared with the PSO algorithm to see the different result provided for each bin, the number of parking stations into each bin is limited because our strategy is applied in the city so the number of parking at a specific area will be limited. The result provided by each simulation describes the total distance of the trip between the current location of the driver toward each nearest parking station inside his area, the total time consumed to find the best trip for each map. So, we will find the best path that exists between the current position of the driver toward each selected parking station. In our case, we assume that a set of initial parameters are predefined by the configuration of the system such as the speed limit which is equal to 45 Km per hour because our system is applied in the city which means that the speed of the vehicle is limited. More parameters are used in our implementation to configure the AS algorithm which is based on a set of local parameters used for controlling the convergence of our program. In Table 1, we can see the rest of the parameters used for our simulation.

Firstly, we compare the performance of both solutions in terms of the total distance of each selected parking slot in the sub-area of the driver. The performance result proves that using our distributed strategy approach, a vehicle can reduce the length of his trip only by following the shortest path proposed by this distributed method (Fig. 9) than the PSO approaches. Each scenario defines a road network combined with a set of parking stations and routes. The distance retrieved from the two approaches represents the total distance of the trip between the start and the destination of the driver. Due to the fact that the distributed approach uses more than one constraint to select the optimal route (such as the traffic data on the route and the linear distance) to propose the best paths and help the driver to reduce the length of his trip to achieve the nearest parking station, it performs better than the PSO algorithm. Simulation feedback, show that the proposed strategy works as well as the PSO algorithm for proposing the optimal path with the minimum travelling distance.

In the next level, we can see that the distributed system aim to reduce the travelling time consumed by the driver based on the AS algorithm by selecting only route with less traffic jam, and based on this mechanism we can reduce the total travelling time consumed by the vehicle to finish his trip. In our case we are based on a set of intelligent agents that implement a probabilistic rule which combine two parameters such as the distance and the traffic data, so, even if one route can be short in terms of distance we should evaluate the traffic jams on this route. We can say that based on this distributed solution we will search for the optimal one in terms of those two constraints (distance, traffic jam). As we can conclude based on the output of the system (Fig. 10), our distributed method reduce the rate of traveling time for all the given scenarios (for each size of bins) compared with the result provided by the PSO solution, because in this last one, we are based only on the linear distance between all the parking stations on each population and we not take into consideration the travelling time required to cross those nodes. So based on those elements, the distributed solution gives us the set of best paths that can be chosen by the driver for each nearest parking station inside the area of this driver.

Finally, we compare the required time consumed to compute and find the optimal solution for the given request by computing the time from handling the request toward the preparation of the best path that will be taken by the driver. The output data, show that our distributed strategy takes advantage of the parallel system to search and prepare the best solutions in a minimum time, also the strategy is based on a set of agents which they work in parallel to reduce the time consumed for arriving at the optimal solution, on the other hand, in the basic strategy (PSO Algorithm) we are based on a sequential process, which consumes more memory and CPU time. Compared with the PSO algorithm (Fig. 11), we can see that our proposed solution is better in terms of the quality of the proposed solution, also our system consumes less time to compute the best solution because of the parallel mechanism in our strategy.

As we can conclude and based on the feedback received from the results data we can see that the introduced method aims to solve the problem based on three levels :

- Searching for finding the best solutions based on the initial constraints (in our case we need to take into consideration the length of the trip and the traffic flow on this route).

- Reducing the time consumed to reach the selected parking stations which are related directly to the state of the traffic on the routes.
- Minimizing the time consumed to find those solutions because in our case we are using a distributed platform based on multiple agents, those agents are running in parallel to find and compute those best solutions, so the parallel behaviour is very important in our case to win mutch time when the system is started.

So the power of this method is represented by the distributed and the parallel process of communication between agents to compute and find the best solutions for this problem, also, in this recent year's several techniques have been made to improve the process of communication between entities which aim to reduce the time consumed for sharing useful data between them. Here, we are using this distributed behaviour to build a new concept of IoT combined with the ACO which will help us to define a new framework that implements the ACO on a distributed platform based on a set of intelligent agents.

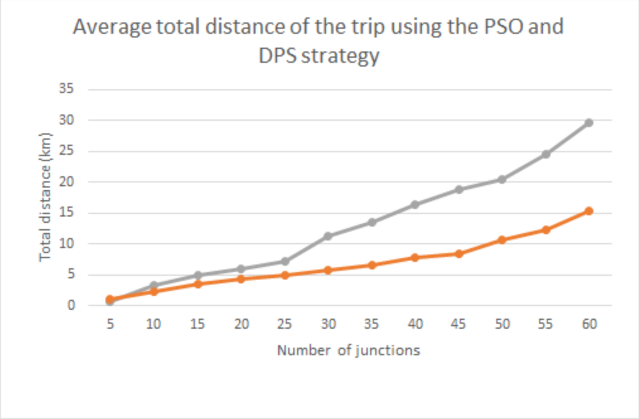


Fig. 9. Average total distance of the trip using the PSO and DPS strategy

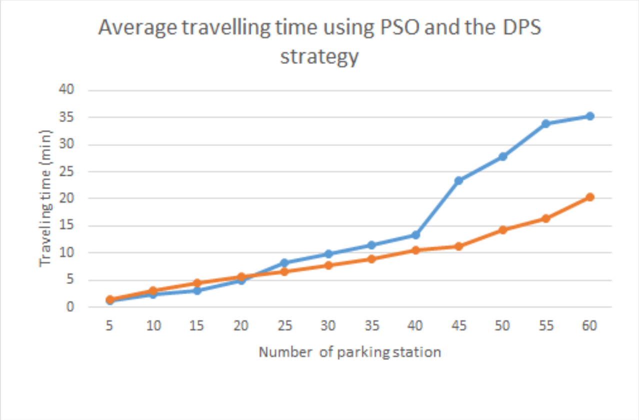


Fig. 10. Average travelling time between PSO and DPS

IX. Conclusion

The concept of building Smart Cities have always been a future vision for humanity and the growth of Cloud technologies and the Internet of Things have given rise to discover new possibilities to improve the services provided by our cities and transfer them to a smart cities. In this paper, we tried to create a new strategy of the parking routing system based on the internet of things combined with a distributed swarm intelligent method to improve the quality of the proposed solutions. At this position, the presented solution aims to help drivers to find the nearest parking places by taken into consideration a set of constraints such as the total travelling time to achieve this parking and the state of traffic road to reach this parking station, so based on these constraints, our solution implements a distributed architecture which contains a set of distributed intelligent agents each one of them has a predefined behaviour. The role of this system is to find the best shortest paths between the location of the driver and the available parking stations. As shown in the result section, we can said that our system will be a useful solution for driver to find the nearest parking slot in a given area and give them the best path planning to reach this parking based on the technique of the ant colony optimization which is dedicated for solving problem attached to the vehicle routing system.

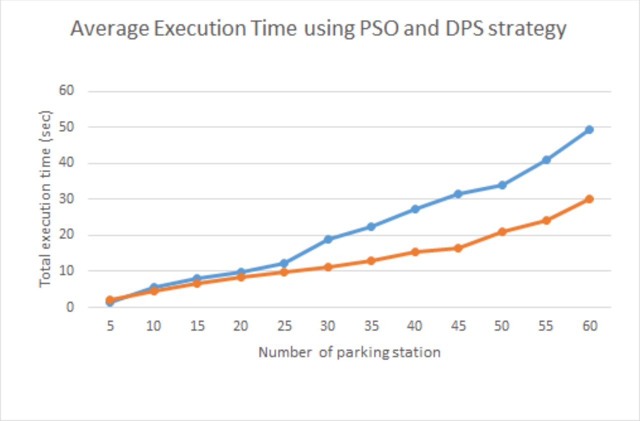


Fig. 11. Comparaison of average execution time to find the optimal solution between PSO and DPS

References

[1] Ameya Naik, Sheetal Vatari, Manjiri Gogate; Real time traffic management using Internet of Things,Tanvi Tushar Thakur ; International Conference on Communication and Signal Processing (ICCSP),2016

[2] Ankit Dubey, Mayuri Lakhani, Shivansh Dave, Jignesh J. Patoliya; Internet of Things based adaptive traffic management system as a part of Intelligent Transportation System (ITS); 2017 International Conference on Soft Computing and its Engineering Applications (icSoftComp)

[3] Aws Naser Jaber et al; A study in data security in cloud computing ;2014 International Conference on Computer, Communications, and Control Technology

[4] B M Mahendra, Savita Sonoli, Nagaraj Bhat, Raju, T Raghu; IoT based sensor enabled smart car parking for advanced driver assistance system; 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)

[5] Guy Krasner , Eyal Katz; Automatic parking identification and vehicle guidance with road awareness; 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)

[6] Hong Zhan, Zhigang Wen , Yuxin Wu , Junwei Zou , Shan Li; A GPS navigation system based on the internet of Things platform;IEEE 2nd International Conference on Software Engineering and Service Science,2011

[7] Hong Guo , Dandan Han , Hongguo Zhang; Using PSO to improve ant colony optimization algorithm;International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things 2014

[8] Ilhan Aydin, Mehmet Karakose, Ebru Karakose; A navigation and reservation based smart parking platform using genetic optimization for smart cities; 2017 5th International Istanbul Smart Grid and Cities Congress and Fair (ICSG)

[9] Jerry Kponyo, Yujun Kung, Enzhan Zhang; Dynamic Travel Path Optimization System Using Ant Colony Optimization;2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation

[10] Jerry John Kponyo, Yujun Kuang, Zejiao Li; Real time status collection and dynamic vehicular traffic control using Ant Colony Optimization;2012 International Conference on Computational Problem-Solving (ICCP)

[11] Kai Lin , Chensi Li, Giancarlo Fortino , Joel J. P. C. Rodrigues; Vehicle Route Selection Based on Game Evolution in Social Internet of Vehicles;IEEE Internet of Things Journal (Volume: 5 , Issue: 4 , Aug. 2018)

[12] Kui-Ting Chen, Yijun Dai, Ke Fan, Takaaki Baba; A particle swarm optimization with adaptive multi-swarm strategy for capacitated vehicle routing problem; 2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom)

[13] Muhamad Asvial , M. Faridz Gita Pandoyo , Ajib Setyo Arifin; Internet of Things Solution for Motorcycle Riders to Overcome Traffic Jam in Jakarta Using EBkSP;2018 International Conference on Information and Communication Technology Convergence (ICTC)

[14] Meijuan Kou, Yanlin Zhao, Hanyu Cai, Xiumei Fan; Study of a Routing Algorithm of Internet of Vehicles Based on Selfishness;2018 IEEE International Conference on Smart Internet of Things (SmartIoT)

[15] M.Elgarej, K.Mansouri, M.Youssfi, N.Benmoussa, H.elfazazi.(2017) ”Distributed Swarm Optimization Modeling for Waste Collection Vehicle Routing Problem”, International Journal of Advanced Computer Science and Applications(ijacsa)

[16] Priti Chakurkar, Sajeeda Shikalgar, Debajyoti Mukhopadhyay; An Internet of Things (IOT) based monitoring system for efficient milk distribution;2017 International Conference on Advances in Computing, Communication and Control (ICAC3)

[17] Pranav Chippalkatti, Ganesh Kadam, Vrushali Ichake; I-SPARK: IoT Based Smart Parking System; 2018 International Conference On Advances in Communication and Computing Technology (ICACCT)

[18] Prabhu Ramaswamy; IoT smart parking system for reducing green house gas emission; 2016 International Conference on Recent Trends in Information Technology (ICRTIT)

[19] Sangita S. Chaudhari, Varsha Y. Bhole; Solid Waste Collection as a Service using IoT-Solution for Smart Cities;2018 International Conference on Smart City and Emerging Technology (ICSCET)

[20] Sokwoo Rhee; Catalyzing the Internet of Things and smart cities: Global City Teams Challenge; 2016 1st International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in partnership with Global City Teams Challenge (GCTC) (SCOPE - GCTC)

[21] Tanmay Chakraborty , Soumya Kanti Datta; Application of swarm intelligence in Internet of Things;2017 IEEE International Symposium on Consumer Electronics (ISCE)

[22] Vasile Gheorghiţă Găitan, Corneliu Octavian Turcu; An internet of things-based distributed intelligent system with self-optimization for controlling traffic-light intersections,Cristina Elena Turcu;International Conference on Applied and Theoretical Electricity (ICATE),2012

[23] Xiaoying Tang , Suzhi Bi , Ying-Jun Angela Zhang; Distributed Routing and Charging Scheduling Optimization for Internet of Electric Vehicles;IEEE Internet of Things Journal (Volume: 6 , Issue: 1 , Feb. 2019)

[24] Xiaobo Zhang, Liangjie Yu, Yong Wang, Guangqing Xue, Yanbo Xu; Intelligent travel and parking guidance system based on Internet of vehicle; 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)

[25] Xing Xiong, Byung-Jae Choi; Design of Genetic Algorithm-based Parking System for an Autonomous Vehicle; International Journal of Fuzzy Logic and Intelligent Systems, vol. 9, no. 4, December 2009 pp. 275-280

[26] Yang Haoxiong , Hu Yang; Congested traffic based on ant colony algorithm for shortest path algorithm; 2015 International Conference on Logistics, Informatics and Service Sciences (LISS)

[27] Yonca Erdem Demirtaş, Erhan Özdemir, Umut Demirtaş; A particle swarm optimization for the dynamic vehicle routing problem;2015 6th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO)

[28] Zhuangkuo Li, Yannan Ma; Particle Swarm Optimization Based on the Average Optimal Information for Vehicle Routing Problem;2013 Sixth International Symposium on Computational Intelligence and Design

An Automated Approach for Identification of Non-Functional Requirements using Word2Vec Model

Muhammad Younas*[1]
Department of Computer Science
Government College University Faisalabad
Allama Iqbal Road Faisalabad, Pakistan 38000

Karzan Wakil[2]
Research Center
Sulaimani Polytechnic University
Sulaimani, Kurdistan Region, 46001, Iraq

Dayang N. A. Jawawi[3], Muhammad Arif Shah[4], Ahmad Mustafa[5]
School of Computing
Faculty of Engineering, Universiti Teknologi Malaysia
Johor Bahru, 81310 Malaysia

***Abstract*—Non-Functional Requirements (NFR) are embedded in functional requirements in requirements specification document. Identification of NFR from the requirement document is a challenging task. Ignorance of NFR identification in early stages of development increase cost and ultimately cause the failure of the system. The aim of this approach is to help the analyst and designers in architect and design of the system by identifying NFR from the requirements document. Several supervised learning-based solutions were reported in the literature. However, for accurate identification of NFR, a significant number of pre-categorized requirements are needed to train supervised text classifiers and system analysts perform the categorization process manually. This study proposed an automated semantic similarity based approach which does not needs pre-categorized requirements for identification of NFR from requirements documents. The approach uses an application of Word2Vec model and popular keywords for identification of NFR. Performance of approach is measured in term of precision-recall and F-measure by applying the approach to PROMISE-NFR dataset. The empirical evidence shows that the automated semi-supervised approach reduces manual human effort in the identification of NFR.**

Keywords*—*Identification; non-functional requirements; semantic similarity; Word2Vec model

I. Introduction

The success of a software system is based on Functional Requirements (FR) and Non-Functional Requirements (NFR). During the requirement elicitation phase, the primary emphasis is on gathering functional requirements, however, NFRs are overlooked. The nature of agile software methodology is also a reason to ignore NFRs [1]. This ignorance of NFRs becomes a cause to produce significant cost issues in software system [2]. U.S. army's computing system for intelligence sharing with troops fighting in Afghanistan and Iraq with a budget of $2.7 billion was rejected due to the issues such as performance, and usability [3]. Electronic Health Record (EHR) was not adopted by the medical community due to lack of usability [4]. A survey revealed that more than 60% of the projects failed due to ignorance of NFR [5].

The agile software development process is based on human-centric requirement engineering which depends on knowledge of stakeholders' regarding application. in SRS documents, software requirements are represented in natural language [6]. The natural languages produce ambiguity and inconsistency in requirement statements. Therefore, it is hard to model and automate the semantic knowledge into requirement engineering activities such as elicitation [7], analysis [6], traceability [8] and reuse [9]. Furthermore, early identification of NFR is critical in the design and architectural concerns of software [10]. Agile software methodologies support a rapid change in software at any stage of development. Due to this rapid change, the importance of NFR identification approach is increased.

The theme of research is about the identification of NFR by manipulating the textual semantic of FR. domain knowledge, also called vocabulary of the domain and NFR knowledge base helps to identify NFR [6]. Furthermore, the automated approach reduces the human or manual effort in the identification of NFR from the requirements document. The scope of research is to identify the non-functional requirements from natural language based SRS documents.

This automatic approach helps requirement engineers and analysts in identification of NFR from the requirement statements in documents, interview notes, memos and reports. The contribution of paper is as follows:
This study

1) extracts the requirement from the document and identify NFR categories,
2) automates and enhance the NFR identification process through semantic similarity measure and pre-processing methods,
3) uses the Stanford Natural Language Parser (NLP) for automaton of identification process,
4) utilizes Wikipedia dump of data for measuring semantic similarity, and
5) enhances the extraction of NFR in terms of increased

precision, recall and F-measure compared to existing work.

Rest of the paper is organized as follows: Section II addressed the background of the study. Section III presents the automated identification approach. Section IV evaluates the approach. Finally, Section V, VI, VII and VIII describe related work, conclusion, limitations and future work, respectively.

II. Background

A. Non-Functional Requirements

Non Functional Requirements are usually known as "ilities"; or as the quality features or attributes of the system [11]. A study considers NFR as systematic requirements [9]. IEEE requirements practices named NFR as constraints. A study surveyed and identified 156 types of NFR and eight (8) categories [9]. The subcategories of NFR are: (i) access control, (ii) audit, (iii) availability, (iv) capacity and performance, (v) legal, (vi) look and feel, (vii) maintainability, (viii) operational, (ix) privacy, (x) recoverability, (xi) reliability, (xii) security, (xiii) usability, and other.

B. Text Embedding Techniques and Semantic Similarity

Word embedding is a process of mapping words and sentences into vector of real numbers. There are different types of methods for mappings such as neural network, co-occurrence matrix, dimensionality reduction, and probability models.

In text classification, the term frequency-inverse document frequency (TF-IDF) is being used for a long time [32]. TF-IDF is a popular term weighting scheme, and more than 80% of text-based recommender systems in digital libraries use TF-IDF [52]. Term frequency means the occurrence of a term in a document. The TF–IDF gets large value both in the given document and in the all documents because it depend on weights, it filters out common terms. ratio is greater and equal to 1 and the value of IDF and (TF–IDF) is greater than or equal to 0. If a term appears in a greater number of times in a document then the IDF and TF–IDF is near to 0.

In order to calculate semantic similarity between words, corpus-based approach named as Latent Semantic Analysis (LSA) is used. LSA uses the distributional hypothesis in which the words having similar meaning are placed in a similar place in corpus [12]. A paragraph is converted into a (m × n) matrix, where m (number of rows) represents the unique words and n represents paragraphs. A mathematical technique named as Singular Value Decomposition (SVD) is applied to reduce the number of rows in the matrix. SVD is used in such a way that the similarity structure in columns does not change.

The Text is converted into vector, the value of cosine of the angle between the two vectors is called the similarity between the two words. The cosine value closer to 1 represents the words are similar and a value closer to 0 means not similar [13]. LSA is very popular and has many strong points, however, there are some limitations. The LSA cannot incorporate the polysemy (multiple meaning of the word). Another limitation is that the LSA does not support Bag of Word (BOW). In BOW, the text is treated as an unordered collection of words [14].

Word2Vec a novel word embedding procedure is designed by Mikolov [15]. The Word2Vec model learns Word2Vector representation using the multi-level neural-network model. The proximity or semantic similarity distance between the words is calculated with the help of Word2Vec model which transforms text into vector [16]. Word2Vect is an example of a group of related neural networks to construct the linguistic context of the word. Each word has a unique vector value. In vector space, the vectors are ranked by considering the context of the words. Each vector is trained in such a way to maximize the probability of the neighbouring word in the corpus as expressed in Equation (1).

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j\in neighb(t)} logp(W_j|W_t) \tag{1}$$

where the list of words is $W = (W_1, W_2, W_3, \ldots, W_t)$, neigb(t) is a set of neighboring word of word W_t and $p(W_j|W_t)$ is the associated word vectors V_{wj} and V_{wt} in hierarchial softmax [15].

Word2Vec is a pre-trained model [17] which is used to find similarity distance between words. Word2Vec model has many advantages on latent semantic analysis [15].

For measuring the performance of similarity measure methods, the similarity distance between word email and words password, color, font, and logon are calculated as shown in Table I. In term of similarity measure, the corpus-based methods (LSA, NGDwiki, and Word2Vecwiki) perform better as compared to the thesaurus-based similarity measure methods (i.e. Wu, Lesk and Resnik) and co-occurrence methods (PMI). In Table I, the similarity word email is calculated with the rest of the words.

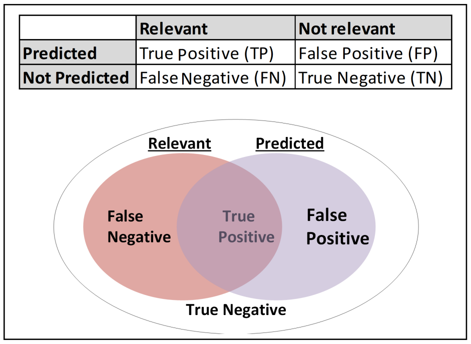
A human with little knowledge about web application would rank the words given in Table I. For example, the words password and logon have more similarity than the word color and font with email. Table I shows that thesaurus-based methods perform almost the same. The method Resnik and Lesk have its similarity value in numbers with minimum zero and maximum infinity. The methods Resnik, Wu, and Lesk have a similarity of email with logon is zero, which shows that these methods do not find the word logon in the WordNet dictionary. In thesaurus-based methods, the common problem is the lack of named entities [18]. This problem arises due to the growth of natural languages data and Expert generated linguistic databases such as WordNet cannot keep up the pace of this growth. Therefore, the selection of such type of similarity measuring methods with WordNet like database might lead to outdated and misleading results.

In the case of co-occurrence-based methods such as PMI, there is a limitation of small corpora. Usually, these approaches rely on available software requirements document. Due to this, their performance is random. The small document is not sufficient for PMI to perform properly. For example, the word email and logon do not occur in the same FR statement in PROMISE dataset. Another study [19] described that the performance of PMI would be enhanced by additional training data. In contrast, the LSA achieve sensible results. However,

TABLE I. SIMILARITY MEASURE OF WORD "EMAIL" WITH DIFFERENT SIMILARITY METHODS

Words	Rensik sim *	Wu sim	Lesk sim*	LSA sim	PMI sim	NGDWiki sim	Word2Vecwiki sim
password	0.78	0.33	16	0.52	0.083	0.49	0.52
color	0.78	0.38	35	0.02	0.083	0.2	0.052
font	0.77	0.3	41	0.01	0.083	0.23	0.1
logon	0.0	0.0	0.0	0.14	0.083	0.355	0.37

* shows that the similarity values are in the range of $(0 - \infty)$ and rest methods in range (0-1)



	Relevant	Not relevant
Predicted	True Positive (TP)	False Positive (FP)
Not Predicted	False Negative (FN)	True Negative (TN)

Fig. 1. Block diagram for precision-recall terms

in their proposed approach the solution is brute force based and LSA has itself a computation intensive solution [18].

Word2Vec embedding has preference over LSA due to linear regularities among words [15]. The Wikipedia used in Word2Vec embedding has better results as shown in Table I. Although the NGD is using Wikipedia, their study shows that there is overhead of training Wikipedia data in converting Wikipedia XML dump into the local database to use it NGD and database size is 26.7 GB. This training is essential because NGD is designed for Google API hit count based. The API provides the service to developers to access and integrate the Google API in their program. Furthermore, there is a limit of quota enforced by Google to hit. Therefore, Word2Vec is used in this method, it has no issue of portability and compatibility with Wikipedia and it produces sensible results. A study [20] used TF-IDF and Word2Vect for classification of text. The study compares the results of both techniques with and without stop words, the results show that without stop word classification performance is better.

C. Evaluation Metrics

To measure the performance of the approach, the study used the precision-recall, and F-measure. The formulas for these measures are as: P = TP / (TP + FP). Here P is precision, a number of corrected or relevant predicted items over a total number of predicted items. The evaluation terms TP, FP, TN and FN are explained through Fig. 1.

In Fig. 1, the term TP means a number of correct predictions, FP means predicted but not relevant item. FN means an item which was relevant but not predicted by the model. The recall measure is defined as R = TP / (TP + FN). The recall is defined as the number of corrected predictions over the total relevant types of requirement sentences. The relationship between the entities can be understood with the help of truth-table and Venn-diagram in Fig. 1.

TABLE II. DESCRIPTION OF PROMISE NFR DATASET

NFR Symbol	NFR	No. of NFR
A	Availability	21
FT	Fault tolerance	10
L	Legal	13
LF	Look and feel	38
MN	Maintainability	17
O	Operational	62
PE	Performance	54
PO	Portability	1
SC	Scalability	21
SE	Security	66
US	Usability	67
F	Functional	255

Furthermore, the relevant mean actual class and predicted mean which is identified by the model. F1 measure means harmonic mean of the precision and recall measures. The formula is F1=2 (P × R) / (P + R). In perspective of NFR extraction, precision and recall is equally important. Recall measure is necessary because requirement engineer tends to identify all NFR. Furthermore, precision of the approach cannot be neglected because large number of false positives results to produce frustration for the requirement engineers.

D. Dataset used for Evaluation

For validation, the NFR identification approach is applied to a PROMISE dataset [21]. The dataset utilized in this study is taken from the Open Science Tera-PROMISE repository. The dataset comprises of 15 requirement specifications of MS student projects of DePaul University. The dataset includes total of 625 requirements written in natural language. The distribution of requirement statements is 255 FR and 370 NFR. Each requirement is classified as FR or NFR. In this dataset the functional requirement is labelled with "F". In the dataset, the NFR has eleven sub-categories along with a count of potential NFR in requirements in the dataset is listed in Table II. The NFR types given in Table II is selected for evaluation of NFR identification approach because several studies [22]–[25] used the PROMISE dataset for their evaluation.

III. RELATED WORK

Software Requirement Specification (SRS) is a document which contains requirement statements usually written in natural languages. In SRS documents, the non-functional requirements are embedded in functional requirements. There are different manual, semi automated and automated approaches to identify NFR from requirement documents [9]. Our approach is close to the automated extraction of NFR from the text documents.

Cleland-Huang, et al. [42] classify NFR in the requirement documents. The study applies TF-IDF with additional variable to specify the frequency of indicator keywords. The study used the collection of weighted indicators keywords to identify NFR. The study observed that the identification of NFR type, “look and feel”, should be improved. The strong point of the approach is that the study provides a base for the automatic extraction of NFR through indicator keywords. The limitation is that the study has very low precision. The recall has a high value of 81%, this is seeming intentional to show performance. A possible reason is that at that time, the high recall was a challenge in the approaches, which the study solved. Due to large false positive, the user gets frustrated in checking the NFR and then discards it.

Zhang, et al. [49] proposed text mining-based technique to identify NFR from requirements documents. The Support Vector Machine (SVM) with linear kernel is used for extraction process and measure the performance of the technique by executing it to the PROMISE dataset. Their study compares the performance of n-gram, single word and Multi Words Expressions (MWE). The comparison is depicted in Table III.

Slankas and Williams [32] present a tool-based approach. The study extracts 14 distinct NFR types from different documents such as installation manuals, requirement specifications, and user manuals. Study selects indicator keywords using probability analysis. For identification of NFR, they use support vector machine, multinomial Naïve Bays and K-nearest neighbors algorithms and compares the results. Furthermore, the approach is tested on different datasets such as CCHIT, PROMISE, iTrust and openEMR.

Mahmoud and Williams [18] present an information theoretic approach for identification of NFR from SRS document. The indicator potential keywords are selected through clustering algorithms. For classification of NFR, the study calculates the semantic similarity between the words using Normalized Google Distance (NGD). For evaluation of approach, The dataset such as SafeDrink, SmartTrip and BlueWallet are used. The average precision of the approach is 53% and recall of 83%.

Another study [50] extracts NFR from the user review of WhatsApp and iBooks. The approach is keyword augmentation based. The keywords are selected using the Word2Vec model. For classification of NFR, the study used supervised learning approaches such as Naive Bayes, J48, and Bagging. In these approaches, the Naïve revealed the best results. Three words to vectorization traditional techniques are used in their study, such that BOW, CHI2, TF-IDF. For enhancing the performance of word to vectorization, they used AUR-BOW, which makes use of Word2Vec to exploit textual semantics to augment user reviews. The strong point of the approach is used keyword augmentation and semantic similarity. The limitation of the study is that the study only considers four NFR (reliability, usability, portability, and performance). Furthermore, they evaluate their approach to self-designed dataset.

Majority of the studies used supervised learning method for detection of NFR. It is observed that supervised learning methods are labor-intensive and have the disbursal of training the model. In case of unavailability of training data, the manual work is required to prepare training data. If the size of the dataset is large, then large training data is required to achieve acceptable results. On the change of the domain of the dataset, the same process is repeated to train the model.

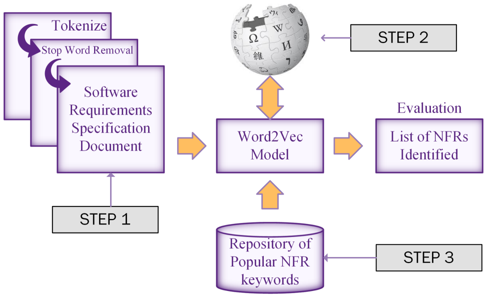


Fig. 2. NFR Identification Procedure

Furthermore, the pre-processing is effective in text mining and machine learning methods. In most of the existing studies, the performance is low in term of precision-recall. There is a need to adopt a semantic similarity-based method to identify the NFR. Internet-based data such as Google data repository and Wikipedia data help in similarity measure.

In Table IV, the existing studies are analyzed with respect to features. The study [42] utilized basic pre-processing and indicator keywords for identification, the Cleland study compares different algorithms. Another study [32] used the synonyms of words and hypernym of potential keywords. The study [18] by Mahmoud and Williams uses three different datasets for evaluation. However, the dataset was not shared due to a confidentiality agreement.

The proposed NFR Identification (NFRId) approach focused on three artifacts to improve the performance. Preprocessing is applied on all requirement sentences before identification process. The literature shows that pre-processing has a key role in improving the performance of machine learning based approaches [51]. Potential indicator keywords are selected by applying probabilistic analysis. The third reason for the improvement is the use of semantic similarity distance for identification. The NFRId has strong point against the existing studies in terms of using semantic similarity measure instead of string matching. Furthermore, NFRId approach has less dependency of trained data as compared to supervised learning based approaches.

IV. Automated Approach for NFR Identification

The study used the three steps procedure to explain the NFR identification (NFRId) approach. The approach utilizes repository of keywords contains the indicator NFR keywords, Word2Vector model is trained with Wikipedia and preprocessing methods for identification of NFR in the requirements document. The steps of identification approach are shown in Fig. 2.

1) Pre-Processing: Non-functional requirements are enclosed in Functional requirements and are usually express in some natural language. Usually, a manual process is adopted for identification of NFR from FR by the requirement engineers. There are some studies that proposed manual, semi automated and automated solutions for identification of NFR in a document [26]–[29]. The pre-processing techniques have a significant role in improving the performance of machine

TABLE III. OVERVIEW OF RECENT STUDIES AND EVALUATION MEASURES

Study	Method / Technique	Dataset	Feature	Advantages	Drawbacks
Cleland-Huang [42]	TF-IDF	PROMISE	The frequency of indicator	Average classification recall is 81%, TF-IDF with extra parameter frequency of indicator term	Large no. of False Positive, precision is 0.12%
Zhang, Yang [49]	SVM-linear kernel	PROMISE dataset	n-gram, individual word, (MWE)	Less effort is required by the analyst in term of preparing training dataset	Consider look and feel, security, legal, usability
Slankas and Williams [32]	KNN, SMO, MNBs, NFRLocator Tool	PROMISE, CCHIT, openEMR, iTrust	Sentence representation (SR), vertex distance logic	Tradition pre-processing, probability analysis in indicator keyword selections	Low value of accuracy 0.38
Mahmood and Williams [18]	NGD, K-mean Clustering	SmartTrip, SafeDrink, BlueWallet	Normalized Google Distance	Use an unsupervised learning approach, semantic similarity based	Evaluated by private dataset not standardized
Lu and Liang [50]	Naive Bayes, J48, and Bagging BOW, CHI^2, TF-IDF	User reviews from iBooks and WhatsApp	Word2Vec	Semantic similarity, keyword Augmentation	Selected NFR, (reliability, usability, portability, and performance), dataset self-prepared

TABLE IV. FEATURE WISE ANALYSIS OF THE EXISTING STUDIES

Study	Word2Vec / NGD	TF-IDF	Semantic similarity	Indicator keywords	Traditional pre-processing	POS tagging / Language Parser	Supervised learning	Public dataset
Cleland-Huang, Settimi [42]	X	✓	✓	✓	✓	X	✓	✓
Zhang, Yang [49]	X	✓	X	✓	✓	X	✓	✓
Slankas and Williams [32]	X	✓	X	✓	✓	✓	✓	✓
Mahmoud and Williams [18]	✓	X	✓	✓	✓	X	X	X
Lu and Liang [50]	✓	✓	✓	X	X	X	X	X
Proposed NFRId approach	✓	X	✓	✓	✓	✓	X	✓

TABLE V. TOKENIZATION, STOP WORDS REMOVAL AND LEGITIMIZATION EXAMPLE

Original statement	The system shall display the Events in a graph by time.
Stop word removal	system shall display events graph time.
Punctuation removal	system shall display events graph time
Lemmatization	system shall display event graph time

learning based approaches such as feature selection, extraction and classification [18], [30]–[32]. The identification approaches usually convert text into vector form, so, there is no meaning of arrangement of words or tokens. Tokenization is a process of breaking text into pieces (words or sentences) called as tokens.

The process of removing unnecessary words is called filtering of data. The unnecessary words are stop-words and punctuation (“the”, “in”, “by”, “of”). The punctuation decreases the ambiguity of meaning. As the words are transformed into vector. So, punctuations has a least significant effect on words embedding’ s order. Therefore, removal of punctuation increases the computation performance of the extraction process [33], [34]. The effect of each process on requirement statement with the example is listed in Table V.

2) *Training of Word2Vector model:* In the second step, Wikipedia dump of data is used to configure Word2Vec model [35]. The size of a dump of Wikipedia article is about 8 GB in compressed form. After converting into plain text, the dump file is of size (down to 13GB) by using the genism script. The plain text is pre-processed and converted into Word2Vec model format, also called word to vector embedding through Genism toolkit. Each topic in Wikipedia is converted into one sentence. This process takes a long time with good CPU speed (take 7+ hours on mac PRO (CPU is 4 core and RAM is 16G) [36]. Pre-trained embedding has an important role to achieve better generalization [37], [38]. Therefore, Word2Vec model is configured on Wikipedia dump of data [39]. For This study, a Genism Word2vec model built on the Engl5ish Wikipedia (Feb 2015), with 1000 dimensions, 10cbow, and no stemming is used [40].

3) *Indicator keywords:* The source of indicator terms is from different studies [32], [41], [42] and quality standards. These popular keywords are normally used to express certain non-functional requirement. The indicator keywords have an important role in the classification of NFR from a text document [41], [43]. The other sources for indicator keywords are quality standards ISO/IEC 25010 [44] standards document and IEEE Standard 610.12 [45].

In experimentation, the study used PROMISE dataset with a total of 625 sentences, 370 of them is annotated as “NFR”, while 255 of them as “FR”, for further use, will be referred to as CorpusNFR and CorpusFR, respectively. Like Chung, et al. [46] approach, the paper measures the probability of potential indicator keywords in the dataset. The frequencies of the words are used for calculating the probability of indicator keywords. Each group is ranked according to the feature probability measures [47]. The probability of the keywords is measured with the following formula.

TABLE VI. List of Indicator Keywords

NFR type	popular keywords
Availability	Availability achieve addition available schedule year period
Legal	Legal law regulations audit license standard custodian definition scope jurisdiction lawyer regulation insurance standard comply ramification liability
Look and feel	access color font graphic green magnify picture red simple blue look feel schema thump appealing
Maintainability	able change configurable integrate maintain new support update
Operational	Operational format mysql infrastructure interoperability machine platform extraction model operate interchange
Performance	date memory processor refresh response startup second speed hour trans transmit time signal live
Scalability	Scalability scalable multiple capable concurrent handle maximum simultaneous
Security	password authenticate authorize protect allow decrepit deny attack malicious protect login email log register role encrypt biometric sensitive restrict prevent
Usability	Usability wrong learn drop realtor voice collision easily successfully estimator intuitive easy enterer word community help symbol training conference let map
Fault tolerance	Fault avoidance tolerance failure unavailable remain restored offline operate remain
Portability	Portability portable

$$
\begin{aligned}
&P(Word) \qquad (2)\\
&= \frac{\#ofWordsinCorpusNFR}{(\#ofWordsinCorpusNFR) + (\#ofWordsinCorpusFR)}\\
&\quad + \frac{\#ofWordsinCorpusNFR}{(\#ofWordsinCorpusFR + 1) \times \alpha}
\end{aligned}
$$

where α is a constant for scaling factor. The smaller value of α, higher the scalability. In this study, the $\alpha = 10$ is used. The keywords selected by Equation (2) are listed in Table VI.

4) Execution of NFR Extraction Approach: The requirement statement has N number of words $W = W_1, W_2, ...W_N$ and each NFR type have M number of variant or morphological words $R = R_1, R_2, ...R_M$ to represent NFR type taken from literature as listed in Table VI. Word2Vec is a model trained on Wikipedia data to calculate the similarity between two words. The procedure for calculating the sentence similarity with a particular NFR type is expressed in the form of the mathematical formula given in Equation (3).

$$SentSim = \frac{1}{N} \sum_{i=1}^{N} \max_{0 \leq j \leq M} Word2Vec_{wikipedia}(W_i|R_i) \quad (3)$$

Equation (3) is implemented in the in the pseudocode given below:

1) Extract the first requirement sentence from the requirements document and get tokens.
2) Find the maximum similarity value between the first token of requirement sentence and all variants of the first NFR type given in Table VI. Repeat the same process with the second token and so on. Find the average value of all the token having the highest similarity value.

TABLE VII. Some Sample NFR Types and a Similarity Value

Requirement	Availability	Legal	Look and feel	Maintainability	Operational	Performance	Scalability	Security	Usability	Fault tolerance	Portability	Result
application color schema forth department homeland security	0.384065	0.248786	0.938915	0.53095	0.49276	0.322887	0.238473	0.346868	0.307832	0.305455	0.314048	LF
system form table system form table	0.235096	0.233892	0.298771	0.33138	0.335732	0.267577	0.238804	0.223263	0.275927	0.262082	0.286431	F

3) Repeat step 2 for all NFR types listed in Table VI and find the average similarity value.
4) The NFR type having the highest average similarity value calculated in step (2) and (3) is our required NFR type.
5) Repeat step (1 to 4) for all sentences in requirements documents one by one.

The pseudo code finds the similarity value between a requirement statement and NFR type. Take the first word of requirement statement and find similarity value of the NFR variant having a maximum value. Take the second word of requirement statement with the highest similarity value and so on. The Algorithm in pseudo code returns the average similarity value of all words in the requirement statement.

In Table VII, the similarity value against each NFR type is calculated by using Equation 3 and the procedure described in the pseudo code. The requirement given in Table VII is in preprocessed form. In the evaluation, only the NFR type with the highest value but greater than threshold similarity value (λ) is considered as the extracted NFR type, otherwise considered as FR. The Threshold value taken in Table VII is λ =0.5. In Table VII, the highlighted value meets the criteria and identified as Look and Feel (LF). Other NFR types having similarity distance values greater than threshold similarity value (λ) but not highest are discarded.

For results, the study executes the approach and get results by applying pre-processing method given in Section III-1. The results are calibrated with the threshold similarity value (λ) as shown in Table IX.

V. Evaluation

Performance is measured in term of precision, recall, and F-measure. The precision-recall is a method explained in Section II earlier. We used the PROMISE NFR [48] data set for the evaluation. The classification breakdown of different NFRs in SRS document is given in Table VIII.

Precision and recall are equally important in extraction process. The precision increases the satisfaction of analyst. In the NFR extraction perspective, the proposed NFRId solution helps the analyst in the extraction of NFR and finally, an analyst re-confirms the NFR types. So, maximum recall also helps the analyst in a greater number of NFRs identifications.

TABLE VIII. Performance of Each NFR Types at Threshold λ=0.5

Requirements	TP	FP	FN	Precision	Recall	F-measure
A	12	18	9	0.4	0.5714	0.4706
L	6	5	7	0.5455	0.4615	0.5
LF	12	6	26	0.6667	0.3158	0.4286
MN	4	23	13	0.1481	0.2353	0.1818
O	27	14	35	0.6585	0.4355	0.5243
PE	18	5	36	0.7826	0.3333	0.4675
SC	7	2	14	0.7778	0.3333	0.4667
SE	39	20	27	0.661	0.5909	0.624
US	39	10	28	0.7959	0.5821	0.6724
FT	4	53	6	0.0702	0.4	0.1194
PO	0	0	1	0.0000	0.0000	0.0000
F	172	129	83	0.5714	0.6745	0.6187
Average				0.5065	0.4111	0.4228

TABLE IX. Performance of NFR Identification Approach at different value of (λ)

Threshold (λ)	Precision	Recall	F-measure
0	0.3372	0.4327	0.3484
0.1	0.3372	0.4327	0.3484
0.2	0.3372	0.4327	0.3484
0.3	0.3375	0.4327	0.3486
0.4	0.4531	0.4449	0.4151
0.5	0.5065	0.4111	0.4228
0.55	0.503	0.3786	0.3988
0.6	0.5039	0.3448	0.3694
0.61	0.5073	0.3442	0.3695
0.65	0.5064	0.321	0.3522
0.7	0.5379	0.2932	0.3356
0.8	0.552	0.2194	0.2563
0.9	0.6225	0.1413	0.1546
0.99	0.034	0.0833	0.0612

Highlighted values show the best value of the measure

Based on the literature, pre-processing has importance in the field of information retrieval methods. So, the study first applies traditional pre-processing on NFR extraction approach. The study applies tokenization, stop word removal and lemmatization before identification process. The highest value of precision, recall, and f-measure at different threshold values of λ (0-0.99) are given in Table VII. In Table VII, the highest value of recall is 45% at λ=0.4, the highest value of F-measure 42% at λ=0.5 and precision 62% at λ=0.9. It is noted that precision value (62%) is at the cost of recall only 14%. The average performance of extraction with respect to maximum F-measure value is as precision of 50%, recall of 41%, and f-measure of 42% at λ=0.5. The performance of identification approach varies on different values of λ. The effect of change is given in Table IX.

Table X shows that the overall performance of the proposed NFRId approach outperforms the Cleland study and Slankas study. All the studies given below in table are used tera-PROMISE NFR dataset. The Slankas study did not describe the performance of their approach in term of precision-recall while evaluating on PROMISE dataset. The automated approach reduced the manual human effort in the identification of NFR from the requirements document.

VI. Conclusion

TABLE X. Comparison of the Proposed Approach with Existing Studies

Study	Precision	Recall	F-measure
Cleland-Huang, et al. [42]	0.147	0.626	0.239
Slankas and Williams [32]	-	-	0.38
Proposed NFRId Approach	0.5065	0.4111	0.4228

In supervised learning, a significant number of pre-categorized requirements are needed to train text classifiers. These pre-categorized requirements are generated by manual classification then the trained model produces better results in terms of identification of NFR. The other limitation of the supervised approach is that if the model is trained for one domain and work well, however, it may not work on some other domain. Experts from different domain use different terminologies. So, you must re-train or re-tune the model. Furthermore, supervised l earning approaches are effective for a small system but face challenges on a large scale or when the systems are not well structured. In this paper, an NFRId approach is proposed and designed for identification of NFR from the requirements document. Our approach does not need pre-categorized requirements for training of Word2Vec model. Therefore, the human's manual effort is reduced in NFR identification process. In the approach, we find the similarity distance between popular NFR keywords and requirements statements. The similarity distance is measured by using Word2Vec model which is pre-trained on Wikipedia dump of text data. The approach is applied on NFR dataset taken from PROMISE dataset repository and performance is measured in term of precision-recall and F-measure. Our result in term of F-measure is 0.47.

VII. Limitations

1) Indicator keywords are focused for PROMISE dataset. so, there is chance to affect the performance for another dataset, for example, medical domain has a different set of word to represent the NFR type.
2) Number of NFR types are bound by the Creator of dataset who labeled it.
3) The dataset has some misconception in labeling of NFR. For example, R2.18 "The product shall allow the user to view previously downloaded search results, CMA reports and appointments" is labeled as NFR in the tera-PROMISE dataset by its creators, however, this is a functional requirement. Our approach also detects it as a functional requirement. The selection of different NFR in the tera-PROMISE dataset seems to some bias.
4) Word2Vec model is bound to Wikipedia vocabulary bank. If the word present in the requirements are not present in Wikipedia, the model cannot find similarity value. It is the limitation of Word2Vec model.

VIII. Future Work

NFRId approach uses unsupervised learning-based model, however, it uses indicator keywords which is a manual process so, as a whole, the approach is semi supervised learning based. In future work, there should be an approach that will use some way to extract indicator terms from clustering or some unsupervised way then our approach becomes a fully unsupervised approach.

Acknowledgment

The authors would like to thank to Universiti Teknologi Malaysia (UTM) for providing resources for this research.

REFERENCES

[1] W. M. Farid, ”The NORMAP Methodology: Lightweight Engineering of Non-functional Requirements for Agile Processes,” in 2012 19th Asia-Pacific Software Engineering Conference, 2012, pp. 322-325.

[2] B. Yin and Z. Jin, ”Extending the problem frames approach for capturing non-functional requirements,” in Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on, 2012, pp. 432-437.

[3] C. Hoskinson, ”Army’s Faulty Computer System Hurts Operations,” Politico, 2011.

[4] J. Bertman, N. Skolnik, and J. Anderson, ”EHrs get a failing grade on usability,” Internal Medicine News, vol. 43, p. 50, 2010.

[5] V. Bajpai and R. P. Gorthi, ”On non-functional requirements: A survey,” in Electrical, Electronics and Computer Science (SCEECS), 2012 IEEE Students’ Conference on, 2012, pp. 1-4.

[6] M. Sadighi, ”Accounting System on Cloud: A Case Study,” in Information Technology: New Generations (ITNG), 2014 11th International Conference on, 2014, pp. 629-632.

[7] L. Xu, G. Tan, X. Zhang, J. Zhou, and Ieee, ”Aclome: Agile Cloud Environment Management Platform,” in 2013 Fourth International Conference on Digital Manufacturing and Automation, ed, 2013, pp. 101-105.

[8] v. one. (2017). 11th annual stat of art agile sruvey Available: http://stateofagile.versionone.com/

[9] N. Kannan. (2012, 2-1-2016). 6 Ways the Cloud Enhances Agile Software Development. Available: ttp://www.cio.com/article/2393022/enterprise-architecture/6-ways-the-cloud-enhances-agile-software-development.html

[10] X. Li, T. Guozhen, Z. Xia, and Z. Jingang, ”Aclome: Agile Cloud Environment Management Platform,” in Digital Manufacturing and Automation (ICDMA), 2013 Fourth International Conference on, 2013, pp. 101-105.

[11] versionone. (2015, 12/01/2015). 7th ANNUAL STATE of AGILE VERSIONONE® Agile Made Easier DEVELOPMENT SURVEY. Available: http://www.versionone.com/pdf/7th-Annual-State-of-Agile-Development-Survey.pdf

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, ”Indexing by latent semantic analysis,” Journal of the American society for information science, vol. 41, p. 391, 1990.

[13] S. T. Dumais, ”Latent semantic analysis,” Annual review of information science and technology, vol. 38, pp. 188-230, 2004.

[14] V. Abedi, M. Yeasin, and R. Zand, ”Empirical study using network of semantically related associations in bridging the knowledge gap,” Journal of translational medicine, vol. 12, p. 324, 2014.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ”Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[16] Y. Goldberg and O. Levy, ”word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method,” arXiv preprint arXiv:1402.3722, 2014.

[17] word2vec. (2013). WordeVec. Available: https://code.google.com/archive/p/word2vec/

[18] A. Mahmoud and G. Williams, ”Detecting, classifying, and tracing non-functional software requirements,” Requirements Engineering, pp. 1-25, 2016.

[19] R. Budiu, C. Royer, and P. Pirolli, ”Modeling information scent: A comparison of LSA, PMI and GLSA similarity measures on common tests and corpora,” in Large scale semantic access to content (text, image, video, and sound), 2007, pp. 314-332.

[20] J. Lilleberg, Y. Zhu, and Y. Zhang, ”Support vector machines and word2vec for text classification with semantic features,” in Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on, 2015, pp. 136-140.

[21] PROMISE, ”PROMISE Software Engineering Repository data set,” 2010, [online]. avaiable: https://terapromise.csc.ncsu.edu/!/#repo/view/head/requirements/nfr.

[22] J. Slankas and L. Williams, ”Automated extraction of non-functional requirements in available documentation,” in Natural Language Analysis in Software Engineering (NaturaLiSE), 2013 1st International Workshop on, 2013, pp. 9-16.

[23] J. Cleland-Huang, R. Settimi, X. Zou, and P. Solc, ”Automated classification of non-functional requirements,” Requirements Engineering, vol. 12, pp. 103-120, 2007. [

[24] D. Sunner and H. Bajaj, ”Classification of Functional and Non-functional Requirements in Agile by Cluster Neuro-Genetic Approach,” International Journal of Software Engineering and Its Applications, vol. 10, pp. 129-138, 2016.

[25] A. Rashwan, O. Ormandjieva, and R. Witte, ”Ontology-based classification of non-functional requirements in software specifications: a new corpus and svm-based classifier,” in Computer Software and Applications Conference (COMPSAC), 2013 IEEE 37th Annual, 2013, pp. 381-386.

[26] N. A. Ernst and J. Mylopoulos, ”On the perception of software quality requirements during the project lifecycle,” in International Working Conference on Requirements Engineering: Foundation for Software Quality, 2010, pp. 143-157.

[27] A. Casamayor, D. Godoy, and M. Campo, ”Identification of non-functional requirements in textual specifications: A semi-supervised learning approach,” Information and Software Technology, vol. 52, pp. 436-445, 2010.

[28] A. Mahmoud, ”An information theoretic approach for extracting and tracing non-functional requirements,” in 2015 IEEE 23rd International Requirements Engineering Conference (RE), 2015, pp. 36-45.

[29] W. M. Farid, ”The Normap methodology: Lightweight engineering of non-functional requirements for agile processes,” in Software Engineering Conference (APSEC), 2012 19th Asia-Pacific, 2012, pp. 322-325.

[30] A. K. Uysal and S. Gunal, ”The impact of preprocessing on text classification,” Information Processing & Management, vol. 50, pp. 104-112, 2014.

[31] G. Feng, J. Guo, B.-Y. Jing, and L. Hao, ”A Bayesian feature selection paradigm for text classification,” Information Processing & Management, vol. 48, pp. 283-302, 2012.

[32] J. Slankas and L. Williams, ”Automated extraction of non-functional requirements in available documentation,” in Natural Language Analysis in Software Engineering (NaturaLiSE), 2013 1st International Workshop on, 2013, pp. 9-16.

[33] H. Saif, M. Fernández, Y. He, and H. Alani, ”On stopwords, filtering and data sparsity for sentiment analysis of twitter,” 2014.

[34] C. Silva and B. Ribeiro, ”The importance of stop word removal on recall values in text categorization,” in Neural Networks, 2003. Proceedings of the International Joint Conference on, 2003, pp. 1661-1666.

[35] wikipedia. (2018). Wikipedia:Database download. Available: https://en.wikipedia.org/wiki/Wikipedia:Database_download

[36] TextMiner. (2015). Training Word2Vec Model on English Wikipedia by Gensim.

[37] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, ”Natural language processing (almost) from scratch,” Journal of Machine Learning Research, vol. 12, pp. 2493-2537, 2011.

[38] D. Chen and C. Manning, ”A fast and accurate dependency parser using neural networks,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 740-750.

[39] Gensim. (2017). training word2vec on full Wikipedia. Available: https://groups.google.com/forum/#!topic/gensim/MJWrDw_IvXw

[40] wiki2vec. (2018). Generating Vectors for DBpedia Entities via Word2Vec and Wikipedia Dumps. Available: https://github.com/idio/wiki2vec

[41] J. Cleland-Huang, R. Settimi, X. Zou, and P. Solc, ”The detection and classification of non-functional requirements with application to early aspects,” in Requirements Engineering, 14th IEEE International Conference, 2006, pp. 39-48.

[42] J. Cleland-Huang, R. Settimi, X. Zou, and P. Solc, ”Automated classification of non-functional requirements,” Requirements Engineering, vol. 12, pp. 103-120, 2007.

[43] L. Rosenhainer, ”Identifying crosscutting concerns in requirements specifications,” in Proceedings of OOPSLA Early Aspects, 2004.

[44] ISO/IEC. (2008, Oct 2016). Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE). Available: https://www.iso.org/obp/ui/fr/#iso:std:iso-iec:25010:ed-1:en

[45] I. Std. (1990). IEEE standard glossary of software engineering terminology. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=159342

[46] L. Chung, B. A. Nixon, E. Yu, and J. Mylopoulos, Non-functional requirements in software engineering vol. 5: Springer Science & Business Media, 2012.

[47] I. Hussain, L. Kosseim, and O. Ormandjieva, ”Using linguistic knowledge to classify non-functional requirements in SRS documents,” in International Conference on Application of Natural Language to Information Systems, 2008, pp. 287-298.

[48] C. Beutenm, #252, ller, S. Bordag, and R. Assadollahi, ”A private living lab for requirements based evaluation,” presented at the Proceedings of the 2013 workshop on Living labs for information retrieval evaluation, San Francisco, California, USA, 2013.

[49] W. Zhang, Y. Yang, Q. Wang, and F. Shu, ”An empirical study on classification of non-functional requirements,” in The Twenty-Third International Conference on Software Engineering and Knowledge Engineering (SEKE 2011), 2011, pp. 190-195.

[50] M. Lu and P. Liang, ”Automatic Classification of Non-Functional Requirements from Augmented App User Reviews,” in Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering, 2017, pp. 344-353.

[51] Hussain, I., Kosseim, L., and Ormandjieva, O. ”sing Linguistic Knowledge to Classify Non-Functional Requirements in Srs Documents”, in Proceeding of the International Conference on Application of Natural Language to Information Systems, 2008, pp. 287-298.

[52] Beel J, Gipp B, Langer S, Breitinger C. paper recommender systems: a literature survey. International Journal on Digital Libraries. 2016 Nov 1;17(4):305-38.

VoIP QoS Analysis over Asterisk and Axon Servers in LAN Environment

Naveed Ali Khan[1], Abdul Sattar Chan[2], Kashif Saleem[3], Zuhaibuddin Bhutto[4], Ayaz Hussain[5]
Eletrical Engineering Dept. Federal Urdu University of Arts, Science and Technology Islamabad, Pakistan[1]
Eletrical Engineering Dept. Sukkur IBA University, Sukkur, Pakistan[2]
Telecom. Dept. Dawood University of Engineering and Technology, Karachi, Pakistan[3]
Department of Computer Systems Engineering, Balochistan University of Engineering & Technology, Khuzdar, Pakistan[4]
Department of Electrical Engineering, Balochistan University of Engineering & Technology, Khuzdar, Pakistan[5]

***Abstract*—Voice over IP (VoIP) is a developing technology and a key factor in both the emerging cyberspace engineering and also an accomplishment to set up its position in the telecom industry. VoIP technology is based on internet technology; where data packets switching system is used rather than circuit switching. Whereas, an analog signal is changed over into the digital signals in the full-duplex transmission. VoIP technology is replaced with the conventional public switched telephone network (PSTN) system due to the high flexibility and low-cost. The purpose of this research work is to deliberate experimental and computational performance in the view of quality of service (QoS) parameters of VoIP over local area network (LAN) network. The VoIP systems implementation is based on two different operating system framework (Linux and Windows), whereas, Linux-based and Windows-based private branch exchanges (PBXs), such as Asterisk (Linux-open source) and Axon (window-close source) are configured, installed and verified. QoS factors (such as packet loss, delay, jitter, etc.) are observed over the Asterisk and Axon PBXs in a LAN domain with the assistance of Paessler switch activity grapher (PRTG) monitoring tool. The validations of results are looked at for QoS parameters crosswise over both PBXs with data load (i.e., file transfer and HTTP traffic) during VoIP calls. The productivity and execution of Axon and Asterisk have been equated and analyzed over experimental based outcomes.**

Keywords*—*VoIP; asterisk; axon; computational based QoS; LAN; packet loss

I. Introduction

Voice-over-IP (VoIP) is a technology of telecom industry that makes the opportunity to totally modify the conventional telephone communication. VoIP demands are growing rapidly since last decade. The primary thought of this transformation was to accomplish the rise efficacy, versatility, and cost-effective telephony. IP is not a protocol of voice communication, as it was intended to transport the IP data traffic. On the other hand, its worldwide servers, computers, and workstations make its traditional platform and logical for telephony communications.

The primary goal of VoIP is to accomplish better implementation and speedy response time. The tailback is moved from physical connection restriction such as higher data capacity and better throughput to the processing elements of the performance. To assess the performance of the framework distinctive methods should be considered. Performance valuations activities are practice to describe the activities of the framework. The study based performance assessments are useful for both existing and future frameworks. These frameworks are commonly used to improve system performance.

The expression VoIP is utilized as a part of IP communication for dealing with the exchange of voice data (counting fax transmission) in digitalized shape as opposed to in the customary circuit-switching method of the public switched telephone network (PSTN). Conventional telephone facilities exist for over 100 years. Customary telephone framework conveys voice data on two-wire frameworks as per global standard. VoIP telephone is still comparatively new idea [1]. VoIP is the innovation that permits IP based systems to be utilized for voice applications, for example, communication, voice texting, and video chatting. VoIP specialist organizations utilize the Web to convey voice signals from their systems to the client [2]. IP is not a convention for telephonic communication since it was intended to carry data information. Nonetheless, it is generally used in PCs, workstations and server make it sensible and advantageous framework for the help of telephonic communication [3]. VoIP is intended to supplant the inheritance of time division multiplexing (TDM) technology and systems with an IP-oriented data system. Digitized voice will be conveyed in IP packets over LAN/WAN systems [4], [5]. VoIP can not convey images effectively. Therefore, VoIP needs the following protocols such as media gateway control protocol (MGCP), real-time protocol (RTP), resource reservation protocol (RSVP), H.323, and numerous others to give other services for VoIP client [2]. Explicitly, VoIP can be characterized as: The capacity to make phone calls (i.e., to do all that we can do today with the PSTN) and to send likeness over IP-based network with an appropriate quality of services (QoS) and a much prevalent cost-effective solution [6], [7].

In VoIP, the voice stream is separated into data packages and after that, it is compacted, and forward to end-user by different paths, contingent upon the most effective ways suggested by the congested network, and so on. At the opposite end, the data packages are reassembled, decompressed, and changed over once again into a voice stream through different devices and programming components, contingent upon the type of the call and end-user [8].

Professional VoIP alike Skype, WhatsAppp, and IMO includes numerous private branch exchanges (PBXs), Internet, WANs and different elements with colossal networks. The scope of the proposed strategy is confined to LAN environment and that excessive examination and investigation VoIP PBXs such as Asterisk and Axon on various Operating Systems. The testing of QoS utilizing overseeing tolls alike Paessler switch activity grapher (PRTG). The exploration work includes:

- Strategy planning and setup of VoIP system by utilizing Linux and Windows Operating Systems with required necessary equipment's.
- Installation and configuration of VoIP PBX on Asterisk (Linux) and Axon (Windows).
- Call establishing along with HTTP and file transfer load over both (Asterisk and Axon) PBXs to evaluate QoS performance.
- Monitoring and analyzing QoS parameters such as Jitter, packet loss, and packet interruption/delay.

In Section 2, we review the strategic role of VoIP in enterprises. Section 3 addresses the Asterisk & Axon designing and implementation. We describe experimentation and analysis in Section 4. In Sections 5, we summaries the findings and conclusion.

II. The Strategic Role of VoIP in Enterprises

A. Value to Enterprises

VoIP is implementing a strategically starring role cleverly; whereas, VoIP applications will revolutionary change business style such as e-commerce, call center, and email as of now have. VoIP is not only a cost-effective method to run telephone calls. It is a radical new scope of voice traffic that would increase the worth for an enterprise. A significant number of VoIP companies are concentrating on the application development for the VoIP market. It will drive a new regime for new business-changing applications. There are unlimited conceivable outcomes, which well expand on its applications that emphases the boosting VoIP market [9]. Deployment of VoIP applications blast internet business.

A case could be voice-based website empowering, whereas online voice contact and dialog for the customer to perusing for purchasing the products [10]. Instead of detaching the PC link; at that point dialing your telephone number to get more product information. The buyer taps on your client service web-symbol, a VoIP link will establish with call agent, who can talk specifically with the client about products. Furthermore, a call agent can guide more website pages or request forms to the client to finalize the deal on the spot.

B. VoIP Flexibility over PSTN

VoIP can give benefits that are harder to apply over the PSTN network. For example:

- The capacity of VoIP to handle more than one caller over a single line connection without any additional phone lines.
- VoIP provides secure calls protocol, for example, secure real-time transport convention (SRTP). The utmost difficult task over the traditional telephone system is establishing a secure call, for example, digital transmission and digitalizing network. A challenging issue in the traditional telephone system is encoded and verified data stream.
- VoIP support location independence system; just a quick and established internet link is expected to get an association from anyplace to a VoIP supplier.

VoIP can integrate and support various services over the internet, including massage services, video streaming and voice calls over file transferring, managing caller list, audio conferencing, etc. [11].

C. Challenges in VoIP

The IP based networks are inalienably less trustworthy. While, the circuit-switch telephony network does not give a secure mechanism that information is conveyed in consecutive order; in other words, it does not support QoS insurances. However, VoIP executions may confront issues extenuating such as jitter and latency [12].

Voice, text, and video are delivered in packets over IP-based network with constant maximal capacity. Such type of framework is tended to DoS assaults and congested than conventional circuit switching networks. Traditional circuit switch network is lacking the capacity to support more connections, but convey the rest of the information without any delay. The quality of real-time data, for example, voice calls over the IP network reduces the performance significantly.

It is difficult to manage static delay because a data packet comes from different paths. Though, Delay can be minimized by identifying delay sensitivity in voice packets such as “DiffServ”. Constant delays are particularly tricky when coming through satellite communication. It occurs due to the long round-trip propagation delay. That is around 400–600 milliseconds over geostationary satellites [13]. A reason for the congestion is delay and packet losses that can be maintained strategic planning of traffic engineering.

IP packets should be properly reassembled at the receiver side. It might be possible, that receiving information have missed, out of order, delayed packets. Though, it is necessary to make sure that the voice streams keep up an appropriate time consistency. The impacts of jitter can be moderated by voice packets storage at jitter buffer on packets receiving and before transforming into analog signals. In spite of the fact that this may cause further delay in the network. This maintains a strategic distance from a condition well known as underrun buffer. Whereas, the received voice packets won't process until the next voice packet received appropriately. Even though, voice packets are delayed or lost in the VoIP network that causes jitter or voice absence during a voice call.

Multi-path routing has been proposed in VoIP communication in which voice packet is received from different routes [14]. It is effective to avoid packet loss and delay. To improve the VoIP call quality, capillary routing has been suggested in which the raptor codes or particularly fountain codes are used for voice data packets diversity to achieve better reliability in the VoIP networks.

D. Susceptibility to Power Failure

IP Telephones and VoIP phone connectors interface with switches or modems router that commonly relies upon the accessibility of mains [15]. Several VoIP vendors offer customer premises equipment (CPE) e.g., routers/modems with batteries based power supplies to guarantee continuous connectivity up to a few hours. These battery-supported gadgets are normally intended for analog phone sets.

The vulnerability of telephone network to electricity failure is a typical issue, even with a customary analog solution in regions where numerous clients buy the latest handset that have various modern features such as wireless connectivity, voice messaging, telephone directory, etc. Modems are currently accessible with lithium batteries that have better battery backup [16].

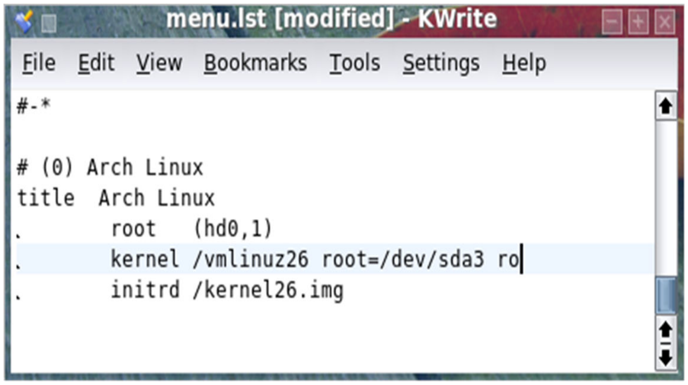
E. Emergency Calls

It is hard to find and arrange clients geologically in an IP based network. Thus, an emergency call can not be routed to a close-by call center. Therefore, sometimes VoIP network may divert the emergency calls to another user. A PSTN call has a straight connection between a phone number and a physical zone [17]. A PSTN phone is connected with telephone exchange through a pair of wires that shows the physical connectivity between user and telephone exchange. Once a line is associated with user number, the phone exchange identifies it with the wires, and this relationship will once in a while change. Even if an emergency call originates from that number, at that point the physical location is known to dialer.

IP based communication isn't so straightforward. A wideband supplier may know the area where the wires ended, but still, it does not really permit the mapping of an IP address to that area. IP numbers are frequently dynamically allotted, so, an internet service provider (ISP) may assign an address for online access, or at the time a wideband switch/router is busy [18]. The ISP perceives individual IP number, yet does not really realize what physical area to which it communicates. The wideband suppliers know the physical area, however, is not really following the IP addresses being used.

III. Asterisk & Axon Designing and Implementation

The Digium is the main investor for the Asterisks evolving, that is an open code software package, that works as soft-PBX for telecommunication. It is based on the Linux framework rather than a windows-based framework. There are numerous Linux distributors, for example, Ubuntu, Gentoo, Fedora Core, and so forth. The issue with fedora core-based Linux operating system is that the kernel system has been altered, therefore, the drivers for the Digium cards can not be accumulated. In addition, as various modules should be integrated with the system, that slow down the whole system performance even with the fast speedy servers. Visibly, utilizing Fedora core is not an ideal choice for the proposed methodology. Whereas, Ubuntu is developed particularly for Intel-based PCs that makes it speedy operating system than the other Linux flavors. Ubuntu-based Linux framework is available at the Ubuntu Linux site [19], plus the details of installation stages for Ubuntu Linux are accessible at the official site [20]. The accompanying is illustrated the imperative steps of Ubuntu Linux installations:



```
#-*

# (0) Arch Linux
title  Arch Linux
.       root   (hd0,1)
.       kernel /vmlinuz26 root=/dev/sda3 ro
.       initrd /kernel26.img
```

Fig. 1. Commands for manuals installation of Linux.

1) **Hard drive segmentation:** Two segmentations are required to effectively installed Linux operating systems. They are root (/) and swap segments. Where root is utilized to store the files systems in the hierarchal order. Whereas, the swap segmentation is utilized for the extra drive space of the Linux framework. Generally speaking, the swap segment quota should be twice bigger than the aggregate of a physical drive. As we can see from the Fig. 1, the root segmentation (˜/dev/sda3) is made with the portion of 145Gb while the swapped segment (˜/dev/sda1) is 2048Mb. Moreover, the boot segment (˜/dev/sda2) is additionally made, despite the fact that not required, keeping the boot loader documents inside the initial 1024 cylinders of the physical hard disk.
2) **Determination of the software installation bundles:** In this section, we pick the software's that we need to install for Asterisk system. It is not important to install each and every bundle since they can be installed later after the completion of operating system installations.
3) **Kernel installation:** The best decision is to choose the kernel v 2.6 which is generally utilized in Linux operating systems.
4) **In the last, the setup of boot loader for Ubuntu Linux:** The GRUB boot loader is chosen by GRUB config-file that will train the GRUB on how to boot the working operating system.

The vmlinuz26 kernel is used. The root segment is placed at ˜/dev/sda3, or in other words, it is placed at the third partition of the physical memory. In addition, the image file of kernel26 carries the commands for GRUB to load the kernel. It is initially used only at the operating system booting time.

A. Asterisk based PBX Implementation

As it earlier said that Asterisk is a VoIP framework, therefore, we will execute VoIP utilizing intense prospect of Asterisk. Asterisk gives the podium to actualizing the VoIP.

To assemble the Asterisk system, we should induce the GCC compiler (3.x. or updated version). GCC is initially composed for the GNU working framework compiler. Asterisk additionally required the bison, a parser-based encoder package which substitutes yacc and ncurses for command-line interface

(CLI) usefulness. The cryptographical archive in Asterisk needs Open shell, plus it's upgrading bundles.

The Asterisk system is taken from the official site [21]. The installation stages that required to run for Asterisk system are as mentioned below:
Stage 1: Decompressed then start from the source code folder:

```
[root˜]#tar -zxvf asterisk-1.4.0-beta3.tar.gz
[root˜]#cd -zxvf asterisk-1.4.0
```

Stage 2: Compose, then install the program's package:

```
[root˜]#./configue
[root˜]#make
[root˜]#make install
```

B. Session Initiation Protocol (SIP) Configuration

Soft-telephones are used for the clients at the LAN system. SIP user accounts are established at the Asterisk system. These records are uploaded at /etc/asterisk/sip.conf folder. There are two different records in this document. One account is represented as AXUSR-1 and the second one is signified as AXUSR-2. The mystery factor is the private key (paswd) that the customer is using to verify with the Asterisk system. The network address translation (NAT) will train the Asterisk system that the customers are located inside the LAN network by using NAT. To deal with the call for AXUSR-1; the AXUSR-1_VOIP dial design setting will be utilized. Whereas, the client AXUSR-2 the dial design setting will be AXUSR-2_VOIP. Whereas, the customer PCs are equipped with X-Lite to manage VoIP calls since it utilizes the SIP-based protocol that is maintained through the Asterisk system. The X-Lite software configuration should match the factors of AXUSR-1 at the sip.conf document of the Asterisk system. Especially, The secret key ought to be announced as the mystery in sip.conf record.

C. Phone Extensions Configuration File

The phone dial design settings account for SIP client and analog phones are placed into the /etc/asterisk/extensions.conf document. The phone dial-design for dial-out is utilized to deal with the calls. For instance, Asterisk will manage and onward the call to SIP client account AXUSR-2. When another distinctive number is squeezed, the call will be sent to client AXUSR-1.

D. Windows-based Axon VoIP Platform

Windows-based IP-PBX Axon has the ability to actualize the versatile PBX answer for the VoIP calls. It can deal with the telephone call on IP based systems. It strengthens boundless expansions for the calls and backings up to 64 phone lines simultaneously. Axon is available on the official Axon site. In the wake of downloading, the distinctive extensions are made for the VoIP system. We have established the two extensions in the current situation. The WAXUSR-1 extensions for client 1 and the WAXUSR-2 extensions will be for the client 2. Presently we have created a call over the extensions with the X-lite software on windows working framework.

TABLE I. Types of Load & Sizes over Asterisk system.

Asterisk	Load Type	Load Size
Experiment 1	File Transfer	800MB
Experiment 2	File Transfer	500MB
Experiment 3	HTTP Load	11759KB
Experiment 4	HTTp Load	2097KB

IV. Experimentation and Analysis

A. Quality of Service (QoS) Parameters Monitored

The accompanying QoS factors are analyzed for both (Axon and Asterisk) PBXs over LAN network with the help of PRTG monitoring tool:

- Packets delay variation
- Maximum jitter
- Packets replicated
- Packets sequences
- Packets losses
- Down-time

B. Paessler Router Traffic Grapher (PRTG)

PRTG is observing software for monitoring, analyzing and mapping the system traffic over the network. It has the ability to analyze and monitor the performance of VoIP data traffic. The ratio of voice calls drop is higher when the user datagram protocol (UDP) packets are not received timely due to the packets disorder and packets losses. PRTG provides the likelihood to act rapidly and keep up a high caliber of administration and furthermore dissect and delineates the distinctive QoS parameters, such as jitter, delay, packets delay, packets loss and so forth.

PRTG gives two approaches to observing and analyzing VoIP execution. we need to include the accompanying sensors modules in the PRTG for analyzing QoS in VoIP.

- Integral QoS estimation sensor
- Cisco IP-SLA sensor

C. QoS Parameters Monitored across Asterisk VoIP Server

To assess the QoS performance distinctive experimentations with different load size and types have been carried by dialing calls over the Asterisk server as shown in Table I. The calls are dial over the customers of the Asterisk servers and after that analyze the QoS factors through PRTG software.

1. File Transfer Loads across Asterisk (800MB & 500MB)

The experiment-1 is carried by dialing a call from the customer AXUSR-1 to customer AXUSR-2 over Asterisk server with the assistance of X-Lite soft-telephone. The steady measurements are observed by the PRTG system for the QoS factors and a call is made during a file transfer over a VoIP network with a fixed load of 800MB. Maximum average jitter is observed around 31ms over the call. The minimum average jitter is found around 2ms. In the event that we consider the entire parametric quantity of jitter factors for each interim of time over the PRTG then we observe that the maximal peak of the jitter stayed 62ms during the call with file transfer load. This 62ms is the pinnacle value as appeared in Table I. The maximal average packet delay is around 236ms and minimal average packets delay is around -117ms. It has likewise been seen thru the experiment that the packets losses are almost 10% for the call time.

Whereas, the experiment-2 is managed with file transfer load 500MB over the VoIP system. Though, the call is produced over the computer for Asterisk server. It evidently recognizes the distinction of QoS parametric factors with various load values. Herein experiment the maximal average jitter is observed around 16ms during call time. The minimal average jitter is observed around 0ms. The maximal packets delay is almost 128ms and minimal average packets delay is -56ms. The packets lost are 0%. So, it has been recognized that the effect of the diverse loads over the system affects the QoS performance.

2. HTTP Loads across Asterisk (11759KB, 2097KB)

Now, experiment-3 is directed with HTTP Load 11759KB by dialing call over the customers of an Asterisk server. It is examined that the maximal average jitter is around 4ms over the call term of time. The minimal average jitter is observed to be 0ms. The maximal average packets delays are 25ms and minimal average packets delays are -22ms. The packets losses are almost 0%.

However, the experiment-4 is implemented with HTTP load 2097KB during the call between the client and the Asterisk server. Though, the maximal average jitter is seen around 6ms. The minimal average jitter is observed around 0ms. The maximal average packets delays are 52ms and minimal average packets delays are around -19ms. The packets losses are almost 1%. The QoS based experimental results are concise in Table II.

D. QoS Parameters Monitored across Axon PBX

In this case, a combination of experiments has been conducted by making calls for windows-based clients over Axon PBX. Furthermore, multiple calls with different load type and magnitudes are dialed between two users and then investigated the QoS factors through PRTG over the Axon PBX as shown in Table IV.

1. File Transfer Loads across Axon(800MB & 500MB)

The experiment-1 is carried by dialing call between the two users (WAXUSR-1 to client WAXUSR-2) over Axon based VoIP-PBX with the assistance of X-Lite soft-telephone. The results are acquired with the help of PRTG soft monitoring tool for the QoS factors by applying constant file transfer load 800MB during the call. The maximal average jitter is estimated around 6ms over the whole call term. The minimal average jitter is observed around 0ms. If we analyze the entire parametric jitter evaluates for each interim of time over the PRTG then we observe that the maximal peak of the jitter stayed 17ms during entire call duration. The minimal average jitter is observed around 0ms. The maximal average packets delay is around 33ms and minimal average packets delay is -30ms. It has likewise been analyzed that the packets loss is 0% for the call span as specified in Table III.

In addition, the experiment-2 is led in a similar way as experiment-1 with 500MB file exchange load of over the network. Therefore, it has been reasoned that the effect of the diverse Loads over the system that affects the QoS esteems. In that experiment, maximal average jitter is observed around 1ms amid the call duration. The minimal average jitter is seen around 0.07ms. The maximal average packets delay is 11ms and minimal average packets delay is -10ms. The packets losses are around 1%.

2. HTTP Loads across Axon(11759KB & 2097KB)

The experiment-3 is conducted with 11759KB HTTP based load. In this experiment, maximal average jitter is around 2ms. The minimal average jitter is observed around 0ms. While, the maximal average packets delay is 10ms and minimal average packets delay variation is -13ms.The packets loss is almost 0%.

While the experiment-4 is applied to with 2097KB HTTP load over the entire call length. Whereas, the maximal average jitter is observed around 1ms. The minimal average jitter is observed around 0ms. The maximal average packets delay is 11ms and minimal average packets delay is -11ms. The packets losses are approximately 0%. The experimental outcomes of QoS over Axon are concise in Table IV.

E. Comparative Study & Analysis of the Results

In the view of the experiments led above for the Asterisk & Axon PBXs to analyze the QoS factors. An analytical investigation is carried out that evidently demonstrate the variation in the QoS performance over the networks.

1. Packets lost in Asterisk and Axon systems with constant file transfer & HTTP loads are presented below in Fig. 2 and 3

A comparison-based analytical study of packets losses over Asterisk and Axon frameworks with file exchange load is demonstrated in Fig. 2. It has been distinctly shown that the packets loss for Axon based systems traces the low limit of percentage as a contrast with the Asterisk call with file transfer load. Thus, the outcome in the terms of packets lost is superior for Axon VoIP framework over the Asterisk system. Moreover, a comparative assessment of both VoIP based PBXs (Asterisk and Axon) frameworks with HTTP load are carried out. Whereas, in Fig. 3, we found that the peak values of packet loss for asterisk reach 10% twice. While it reaches 10% for Axon only once during the entire call duration. Its shows that Axon has better performance than Asterisk with HTTP load.

2. Jitter performance with File Transfer & HTTP Loads for both Asterisk and Axon systems is shown in Fig. 4 & 5

Performance analysis of jitter in Asterisk and Axon framework with constant file load transfer is presented in Fig. 4. It has been evidently revealed from the above graph that the Axon system has less jitter than Asterisk with File Transfer Load. Hence, the outcomes in the term of jitter have a superior

TABLE II. QoS VALUES MONITORED BY PRTG OVER ASTERISK.

Asterisk	Jitter Min	Jitter Average	Jitter Max	Packet Loss	PDV Min	PDV Average	PDV Max
Experiment 1	2ms	8ms	31ms	10%	117ms	0ms	236ms
Experiment 2	0ms	5ms	16ms	0%	56ms	0ms	128ms
Experiment 3	0ms	5ms	4ms	0%	22ms	0ms	25ms
Experiment 4	0ms	3ms	6ms	1%	19ms	0ms	52ms

TABLE III. TYPES OF LOAD AND SIZES OVER AXON SYSTEM.

Axon	Load Type	Load Size
Experiment 1	File Transfer	800MB
Experiment 2	File Transfer	500MB
Experiment 3	HTTP Load	11759KB
Experiment 4	HTTP Load	2097KB

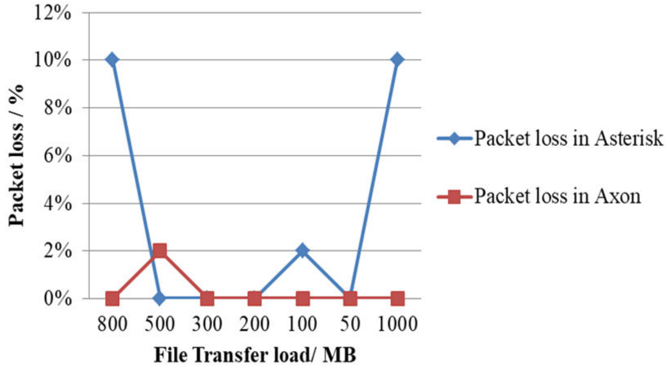


Fig. 2. Packet losses in the Asterisk and Axon with file transfer load.

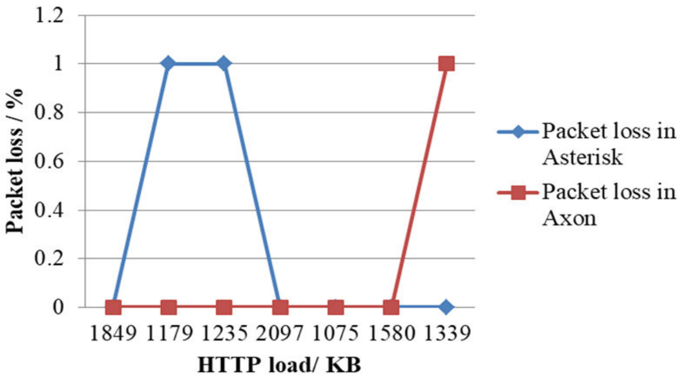


Fig. 3. Packet loss in Asterisk and Axon with HTTP Load

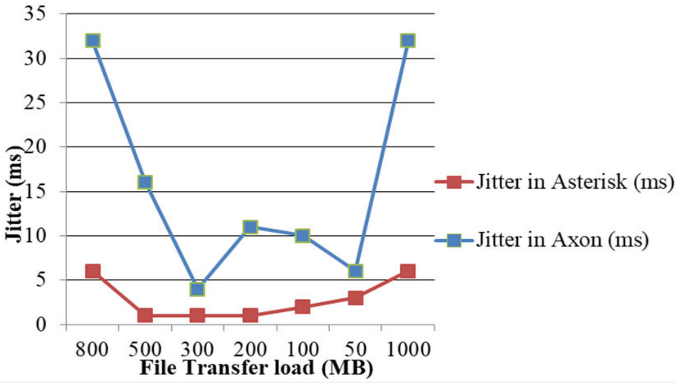


Fig. 4. Average Jitter in Asterisk and Axon with file transfer load

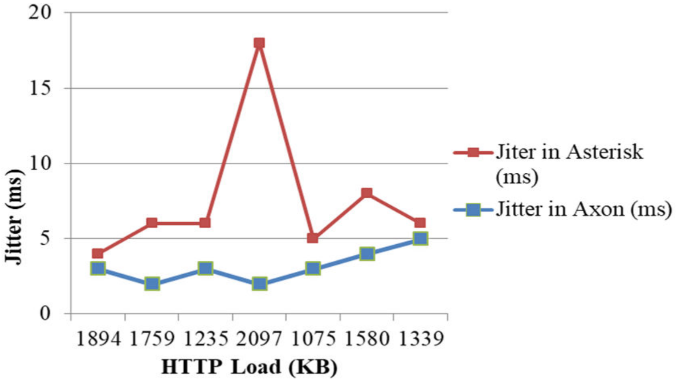


Fig. 5. Jitter in Asterisk and Axon with HTTP Load

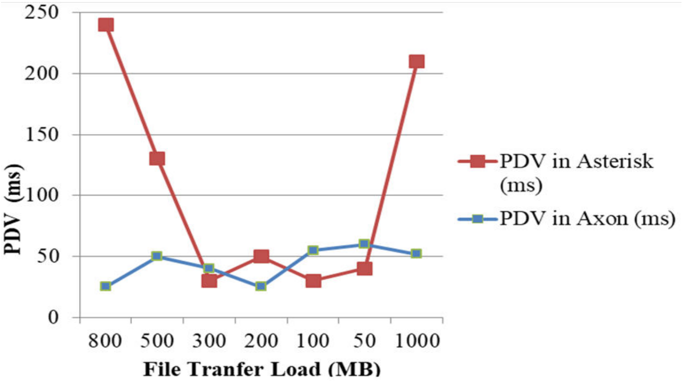


Fig. 6. Packets delay variation in Asterisk and Axon with file transfer load

improvement for Axon network than the Asterisks systems. On the other hand, an assessment is made for jitter performance over Asterisk and Axon systems with HTTP load as shown in Fig. 5. It is evidently depicted that the jitter maximal has better values for Asterisk as compare to the Axon system with HTTP loads. The jitter for Asterisk traces to the value of 17ms for an HTTP Load of 2097KB.

3. Packet Delay Variation with File Transfer & HTTP Loads for both Asterisk and Axon systems is shown in Fig. 6 & 7.

Packet Delay Variation performance of Asterisk and Axon systems with file transfer load is shown in Fig. 6. It has been manifestly shown from the above graph that the Axon system has low-level spikes than Asterisk with File Transfer Load. Therefore, the performance in terms of packet delay deviation is superior for Axon than the Asterisk network. Moreover, valuation is conducted for Packet Delay Variation over Asterisk

TABLE IV. QoS VALUES MONITORED BY PRTG OVER AXON.

Axon	Jitter Min	Jitter Average	Jitter Max	Packet Loss	PDV Min	PDV Average	PDV Max
Experiment 1	0ms	2ms	61ms	0%	30ms	0ms	336ms
Experiment 2	0.07ms	1ms	16ms	1%	10ms	0ms	11ms
Experiment 3	0ms	1ms	2ms	0%	13ms	0ms	10ms
Experiment 4	0ms	1ms	1ms	0%	11ms	0ms	11ms

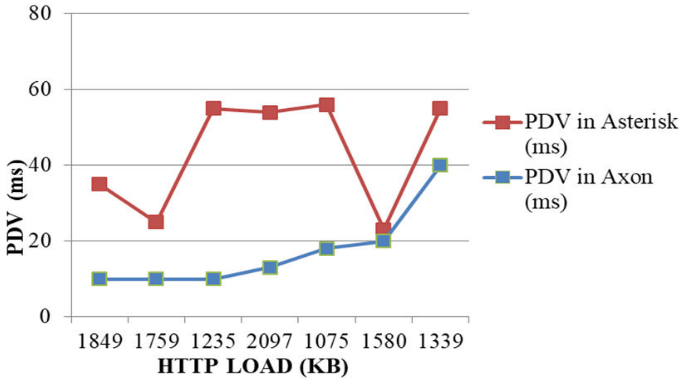


Fig. 7. Packets delay variation in Asterisk and Axon with HTTP load

and Axon systems with HTTP load as shown in Fig. 7. It is basically described that the Packet Delay Variation with HTTP Loads with greater values for Asterisk VoIP system as equating to the Axon system. It's proving that Axon has better performance than the Asterisk system in terms of Packet Delay Variation. Finally, a comparative analytical study described that the Axon based PBXs systems have better outcomes in terms of packet loss, jitter and packet delay variation.

V. Findings and Conclusion

The diverse IP PBXs gives more productive methods for calls steering than the PSTN PBX and with the expanding significance of the IP system. The demand for soft-digitalized PBXs will be expanded. Large numbers of IP based PBXs are available in the market for the call managing. They are giving a decent platform for VoIP networks.

We have analyzed the two versatile PBXs in this research paper. The performance is evaluated in the view of QoS parameters for the VoIP data traffic by producing several experimentations and call processes over a LAN network. As we distinguish that the LAN has an important role in the computer communication system, therefore it has been the emphasis of the overall project to study the VoIP QoS factors over the LAN network.

Several experiments are conducted with a different type (file transfer and HTTP load) and size (800MB, 500MB, 11759KB, and 2097KB) of data load during call over the two different soft PBXs (Asterisk and Axon). It has been summarized that the VoIP network performance of windows-based Axon is vastly improved in the terms of QoS factors as a contrast with the Asterisk VoIP framework in the LAN network.

References

[1] R. Wang and X. Hu, *"VoIP development in China"*, Computer., vol. 37, no. 9, pp. 30–37, Sept. 2004.

[2] B. Goode, *"Voice over Internet protocol (VoIP)"*, proc. IEEE, Sept. 1495–1517, 2002.

[3] S. Micheal, J. Timothy, and J. Micheal, *"Voice over internet protocol (VoIP) traffic management system and method"*, U.S. Patent No. 8,798,039. 5 Aug. 2014.

[4] B. Raj, *"Automatic and seamless vertical roaming between wireless local area network (WLAN) and wireless wide area network (WWAN) while maintaining an active voice or streaming data connection: systems, methods, and program products"*, U.S. Patent No. 7,039,027. 2 May 2006..

[5] W. Kim, T. Song, T. Kim, H. Park, and S. Pack, *"VoIP Capacity Analysis in Full-Duplex WLANs"*, IEEE Transactions on Vehicular Technology., vol. 66, no. 12, pp. 11419–11424, Dec. 2017.

[6] X. Chen, *et al.*, *"Survey on QoS management of VoIP"*, Int'l Conf. Computer Networks and Mobile Computing, Shanghai, China, Oct. 2003.

[7] R. S. Iborra, M. D. Cano, J. H. Haro, *"Performance Evaluation of BATMAN Routing Protocol for VoIP Services: A QoE Perspective*, IEEE Transactions on Wireless Communications, vol. 13, no. 9, pp. 4947-4958, Sept. 2014.

[8] D. Rango, *et al.*, *"Overview on VoIP: Subjective and objective measurement methods"*, International Journal of Computer Science and Network Security, vol. 6, no, 1, pp. 140–153, ——.

[9] C. H. Liao, and P. H. Tseng, *"Influential factors of VoIP adoption of top 500 export-import enterprises in Taiwan"*, Contemporary Management Research, vol. 6, no. 1, pp. ——, 2010.

[10] Keromytis and D. Angelos, *"A comprehensive survey of voice over IP security research"*, IEEE Communications Surveys & Tutorials, vol. 14, no. 2, pp. 514–537, 2012.

[11] Shaw, Urjashee, and B. Sharma, *"A Survey Paper on Voice over Internet Protocol (VOIP)"*, International Journal of Computer Applications, vol. 139, no. 2, 2016.

[12] Ottur, Deepak, *et al.*, *"Method and apparatus for testing in a communication network"*, U.S. Patent No. 9,769,237. 19 Sep. 2017.

[13] H. P. Singh, *et al.*, *"VoIP: State of art for global connectivity: A critical review"*, Journal of Network and Computer Applications, vol. 37, pp. 365–379, 2014.

[14] Serrano, Salvatore, *et al.*, *"VoIP traffic in wireless mesh networks: a MOS-based routing scheme"*, Wireless Communications and Mobile Computing, vol. 16, no. 10, pp. 1192–1208, 2016.

[15] Mushtaq, M. Sajid, *et al.*, *"QoEQoS-aware LTE downlink scheduler for VoIP with power saving"*, Journal of Network and Computer Applications, vol. 15, pp. 29–46, 2015.

[16] Emara, Z. Tamer, I. Ahmed, Saleh, and H. Arafat, *"Power saving mechanism for VoIP services over WiMAX systems"*, Wireless networks, vol. 20, no. 5, pp. 975–985, 2014.

[17] Oorni, Risto, and A. Goulart, *"In-Vehicle Emergency Call Services: eCall and Beyond"*, IEEE Communications Magazine, vol. 55, no. 1, pp. 159–165, 2017.

[18] Manso, Marco, *et al.*, *"The application of telematics and smart devices in emergencies"*, Integration, Interconnection, and Interoperability of IoT Systems, Springer, Cham, pp. 169–197, 2018.

[19] https://www.ubuntu.com/

[20] Sobel, G. mark, *"The application of telematics and smart devices in emergencies"*, A practical guide to Ubuntu Linux, Pearson Education, 2014.

[21] https://www.asterisk.org/

Privacy Preserving Data Mining Approach for IoT based WSN in Smart City

Ahmed M. Khedr[1]
Computer Science Department,
University of Sharjah, Sharjah 27272, UAE

Walid Osamy[2]
Faculty of Computers and Artificial Intelligence,
Benha University, Benha, Egypt.
Qassim University, Buridah, KSA

Ahmed Salim[3]
Math. Department, Zagazig University,
Zagazig, Egypt.
Qassim University, Buridah, KSA

Abdel-Aziz Salem[4]
Department of Basic Science, Faculty of Computers
and Informatics, Suez Canal University.
Qassim University, Buridah, KSA

Abstract*—Wireless Sensor Network (WSN) is one of the most fundamental technologies of Internet of Things (IoT). Various IoT devices are connected to the internet by making use of WSN composed of different sensor nodes and actuators, where these sensor nodes collaborate and accomplish their tasks dynamically. The main objective of deploying WSN-based applications is to make high precision real-time observations, and it is extremely challenging because of the limited computing power of the sensors operating under constrained environments, resource constraints like energy, computation speed, bandwidth and memory, huge volume of high speed, heterogeneous and fast-changing WSN data. These challenges encouraged the researchers to concentrate deeper on exploring data mining techniques to extract the required information from the fast-changing sensor data in WSN and thereby efficiently handle the massive data generated by the WSNs. The increasing need of data mining techniques for WSN has inspired us to propose a distributed data mining technique that effectively handles the data generated by the nodes in the WSN and prolongs the lifespan of the network. Our work provides a novel cluster based scheme to mine the sensors data without moving it to cluster head (CH) or base station (BS) to achieve maximum performance in a WSN environment. The basic idea of the proposed work is that local computations are performed by utilizing the computing power at each sensor node and then the minimum higher level statistical summaries are exchanged, which decreases the energy dissipation in communication as the amount of the sensor data transferred is considerably reduced, and thereby the sensor network lifetime is maximized and also preserve the privacy of the sensor data.

Keywords*—*Distributed cluster-based algorithm; association rules; Internet of Things (IoT); privacy preserving; vertically and horizontally distributed databases; wireless sensor networks (WSN)

I. Introduction

Knowledge discovery from sensor data is an emerging research area and mining data efficiently and effectively from the resource constraint sensor nodes is really challenging. Recently, the researchers on data mining are strongly motivated by the challenges raised by the huge volume, high speed, heterogeneous, rapid and fast-flowing data generated by WSNs and gathering crucial information from the device data has become a topic of active research. As a result, new data mining techniques has been developed and some of the existing approaches has been modified to devise new analytic methods for the massive quantity of data generated from sensor networks. Distinct data mining techniques that deal with extraction and analysis of WSNs data concentrated on clustering [1], association rules [2], [3], frequent patterns [4], [5], sequential patterns [6] and classification [7] have been efficiently applied on sensor data to make intelligent WSN applications. Most of the traditional mining techniques perform intensive centralized computation and is expensive. Moreover, the deployment and implementation of WSNs create a lot of research challenges making the direct application of traditional mining techniques inappropriate to WSN. The huge volume and high cost of storage make it quite impossible to store the fast-flowing WSN data or to inspect it several times. The nature and characteristics of the sensor data, limited communication and computation capabilities of the sensor, and special design and deployment limitations of the WSNs make the application of primitive data mining procedures challenging. Therefore, it is highly required to devise data mining techniques that are capable of handling continuous, fast-changing and extensive data streams of WSN with high dimensionality and distributed nature, and to analyze and process it in single-pass, multi-level, multi-dimensional or online methods of data mining.

With the rapid advancement of sensor network technology, WSN based systems are becoming more popular and are increasingly finding its applications in different areas of knowledge [8]. This resulted in the generation of diverse WSN applications yielding extensive heterogeneous distributed data to be analyzed efficiently and effectively [9]. Such applications are mostly critical and require real-time control and reliable operation as crucial demands.

WSN acts as a virtual layer and has become an intrinsic part of IoT in a secure manner. But to do so, it has to overcome various challenges such as security, integration issues, energy optimization, network lifetime, etc. WSN is like the eyes and ears of the IoT. It is the link which joins the real world to the digital world. And it is also responsible for passing on the sensed real world values to the Internet. IoT based wireless sensor networks (IoT-WSNs) have a wide range of applications in various fields, which allows inter connection of different objects and nodes through Internet. IoT-WSN can be described as collection of enormous sensor nodes deployed with a number of moving objects or devices (such as intelligent cars) over a large area (Smart City) to sense and accumulate various data from the environment and systems for different applications such as weather monitoring, animal tracking, disaster management, bio-medical applications.

It has been proved that a database oriented approach is helpful to manage the dynamic nature of WSN data for those applications [10], [11] and hence it motivated the researchers to treat WSN as a distributed database. Accordingly, WSN can be modelled as a distributed database where sensor devices act as data sources and stores the data with the sensor in the form of database across the network in distributed manner [12], [13]. The main objective of distributed database management [14] on WSNs is to facilitate the energy efficient analysis of massive data sensed by the sensor nodes. In order to extract the data, a WSN database should provide support of robust queries eg.: SQL-like abstractions for interaction with the network [15].

Upon research on sensor hardware, it has been proved that the depletion of energy in WSN [16] is mainly during the exchange of data among the sensor nodes. Different data reduction techniques [17], [18] such as packet merging, aggregation [19], approximation based techniques [21]], data compression techniques [20], [23], [24], [25], [26], [27], [28] and data fusion [22] techniques are being used to handle this problem.

Privacy preserving data aggregation is a challenge in IoT-WSN as it could be spied. The privacy preserving data aggregation has intent to save individual privacy of the nodes using transmission directions, in such a way that the enemies cannot obtain the sensitive information of a specific node. A city is considered smart city when it functions in a sustainable and intelligent way through the integration of all its infrastructures and services by deploying intelligent devices and networks for monitoring and control. Smart city applications such as intelligent transportation systems as well as monitoring public infrastructures (e.g. bridges, roads and buildings) are based on the data aggregation by static and mobile sensors deployed in very large numbers. We handle the data' privacy node at aggregation time by considering that privacy preservation can be achieved if nodes have devices that may be heterogeneous and have different places as a result every node only sense a small part of the whole collected data; their results have to be aggregate to give a whole and global scenario (IoT-WSN will be considered as big distributed database system with vertical partitioning data). This provides fast query processing and preserving the individual sites private data.

We propose a new version of association rule algorithm to work with distributed IoT based WSNs data. This paper provides a methodology in which each CH is represented by an agent, with the potential to decompose global computations into local ones. That means, the computation at CH is executed by exchanging minimal statistical summaries with other agents at sensor nodes and the results of every CH will be globally aggregated at the BS. The main objective of using this distributed approach is to lessen the data transfer and energy utilization of sensors upon exchanging data with central server, which conserves energy and prolongs the lifespan of network to a great extent.

The rest of this paper is divided and structured as follows: The related research of the proposed problem is described in Section II, while integration methodology of the databases is given in Section III. Section IV presents the proposed scheme. The simulation results and comparison with other baseline algorithms are included in Section V and finally, conclusion of the paper is given in Section VI.

II. Related Research

WSNs producing huge volume of data is offering a promising prospect for data analytic and mining techniques to involve in extracting useful information for a wide variety of applications. Existing data mining techniques adopted for WSN are classified depending on its application: whether it is on network side or central side. Sequential mining, frequent mining, clustering and classification are the four classes of data mining approaches commonly adopted in WSNs which use both centralized and distributed approaches. The mining techniques based on these above mentioned classes form the first type of classification. Most of the techniques based on sequential and frequent pattern mining are adaptation of the primitive techniques such as the FP-Growth Algorithm and the influential algorithm like Apriori for association rule determination and learning for extracting potentially valuable information from large amounts of WSN data, while the techniques based on clustering have adapted the data correlation-based clustering, k-means and the hierarchical clustering approaches. Approaches based on classification have adapted several major kinds of classification techniques including k-nearest neighbor classifier, Bayesian networks, Decision tree, Neural Networks and Support Vector Machines, according to the type of classification model adopted. The second type of classification is based on how the data is processed and analyzed- either centrally or in distributed way. The limited computing power of the sensors operating under constrained environments, resource constraints like energy, computation speed, bandwidth and memory, huge volume of high speed, heterogeneous and fast-changing data generated by WSNs, make distributed processing a more preferable solution. The distributed data computation approaches perform mining and other processing of data at sensors locally and then accumulate the results. The third type classification is based on the attitude concerned with solving a specific problem. This level of classification is mainly focused on WSNs performance issues and application issues. The characteristics of the sensor data, precision, accuracy and real-time decision making of the WSN applications often require abundant use of communication and energy. The algorithm proposed in this paper is a distributed frequent pattern mining technique with reduced energy consumption and improved WSN lifetime.

According to the above classification of mining algorithms, we review the existing literature as follows. Frequent pattern mining technique finds the stream of data that frequently occurs in the dataset, with the objective to determine crucial relations existing between the sensors data. Primitive algorithms for frequent pattern mining [29], [30], [31], [32], [33], [34], [39] are resource intensive and cannot be directly used for mining the fast-flowing WSN data. Numerous algorithms are designed for solving the application-based issues of WSNs such as [35], [36], [37], [49]. In this context, the authors of [35] introduced a centralized technique named Data Stream

Association Rule Mining (DSARM) to determine the missed values in sensor data readings due to corruption or loss of messages. It identifies sensors that repeatedly report the same data and estimates the missed values in sensor data by making use of the data communicated by its related sensors. Closed item-sets-based Association Rule Mining (CARM) is a data estimation technique [36], [37] for deriving the recent sensor association rules based on the latest closed item-sets in the sliding window. Other than these, a number of centralized algorithms were also designed to maximize the performance of WSNs [38], [40], [41]. The authors of [38] proposed online one-pass methodology, in which the WSN data-stream is converted to interval list (IL), for inter-stream association rule mining from bulk sensor data stream. A rule-learning model is proposed in [40] to derive powerful rules from sensor data readings, to control and coordinate WSN operations. The sensor pattern tree proposed in [41] is a tree-based data structure for deriving association rules from sensor data by scanning the database only once.

Several distributed approaches exist in the literature for solving application based issues of WSNs and/or maximizing the performance of WSN [43], [2], [42]. Author in [43] presents a distributed method with some spatial/temporal properties to detect frequent event patterns. Their work describes an in-network approach of data mining in which user can specify the spatial and/or temporal proximity for patterns among events in which he is interested in. The sensors collect events accordingly and execute a data mining procedure to identify the pattern among these collected events that satisfies the user specified parameters. The frequent patterns obtained after mining are then converted into association rules. Every node in the network will then send the discovered local patterns to the sink for performing secondary mining on these patterns to create a global picture of the entire network with respect to time and space. However, the communication overhead of event collection and memory consumption of item-set discovery algorithm are the major issues with this methodology. A distributed data extraction methodology is presented in [2] to accumulate the data on sensor in an attempt to lessen the number of message exchanges. Each node in the network is equipped with a buffer and has corresponding entry for support value. Moreover, parameters like time-slot size, support, and historic period are distributed across the WSN. The sensor node will examine the received messages per time slot and sets its buffer entry accordingly. Every node checks its buffer upon completion of historic period and compares the set value with the initially provided support value. The message will be transferred to the sink if the set value is larger or equal to the support value. The potential drawbacks of this technique include delay in critical messages for high support value and the node buffer cost. The authors of [42] proposed a new representational structure named Positional Lexicographic Tree (PLT). Using the specified sensor allocation rules for the event detecting sensors, it stores the sensor's event-detecting status and is a promising structure that can be used for indexing and compressing the sensor data residing in any transactional database. It also facilitates subset checking using data summaries. The way in which the PLT identifies the conditional structure is found to be comparatively easier than FP-tree method. PLT also allows the mining process to be partitioned into various tasks; each of which can be accomplished separately. The issue with PLT is that it requires multiple scans and PLT updates, restricting the effective use of this technique in dealing with WSN data.

Our proposed approach is completely different from the above and to the best of our knowledge this is a new approach to handle the mining problem in WSNs. The key idea of our distributed approach is to treat the whole network as a distributed database where, the sensed data is kept at the sensor nodes, stored in rows, and the columns represent sensor attributes. We do not move the sensor data to the CH or BS. We consider the global data D in implicit format, i.e., the tuples belonging to sensor database D is distributed over all the nodes in the WSN, with each node generating and accumulating its data. In our approach the sensor nodes in the network are grouped into various clusters, each having a CH managing the cluster. The queries generated by the CH will be periodically answered by the CMs belonging to it, based on their readings. Only the statistical summaries will be transmitted to CHs. It accumulates the received summaries and transfers the aggregated computation output to the BS for final results. Rather than traditional queries which mainly focus on the current state of a database, these queries (SQL-like abstractions) are often continuous so that the application will be notified continually about the changes recorded by the sensors [12]. A sensor node sends its response to the current query only if its reading is different from the last recorded reading. This method of aggregation efficiently decreases the huge volume of data transmitted through the network, reducing the computation burden enforced on the BS, which in turn raises the lifespan of the WSN. Furthermore, for efficient execution of query with low energy dissipation and delay, an effective load balancing policy can be adopted in terms of remaining power and the load of the nodes.

III. Databases Integration

As discussed before, each sensor node is assumed to have a relational database with a number of attributes. In this section, we describe different scenarios of databases and the proposed methodology to handle these databases.

A. Nature of Data Distribution

Vertically Distributed Databases: The implicit database D exists as a set of distributed fragments across all the devices in the WSN. Each database component D_i stored at device node (s_i) consists of a set of records (tuples), where every record stores a diverse attribute set. Another component D_j stored at device node s_j $j \neq i$ could contain some attributes shared in common with D_i, as well as distinctive attributes that aren't shared with other components. The databases following vertical strategy of distribution need "Join" operation on all explicit D_i's to build the implicit database D. The planned algorithm will make use of the shared attributes in performing data processing. This approach demonstrates a more practical scenario than using a single key, non-overlapping set of attributes for the components distributed across all the devices in the WSN. Our aim is to enable the participation of these independently designed local component databases to have arbitrary overlapping set of attributes upon collaboration with each other resulting in single global database. The association rule is performed on the database resulting from joining the relations on the sensor devices distributed across the network. *Horizontally Distributed Databases:* The implicit database D exists as a set of distributed fragments $D_1, D_2, \ldots, D_n$ across the sensor nodes $s_1, s_2, \ldots, s_n$ such that every D_i contains unique set of tuples having identical attributes set. The set of tuples present in the distributed D_i's will then be merged together to form the global implicit relation D. The proposed methodology is applicable on any of the above mentioned distribution of database, either horizontal or vertical.

B. Problem Formulation

The entire WSN is treated as a distributed database such that each sensor node has a local component database following either horizontal or vertical data distribution. The join of these databases at the BS or end user creates the global implicit database D containing significant data for computation and mining tasks. The key objective of the proposed algorithm is to reduce the data exchanges by retaining the local data with the node itself. Only the local results of D_i of CMs will be transmitted to be aggregated at their CH and then the aggregated results of CHs will be transmitted to be aggregated at BS. Let A be the attribute set of the global implicit database D and A_i be the attribute set of the local database D_i at sensor node s_i. Then A can be defined as the union of all the attributes between all the local databases within the network.

$$A = \bigcup_{i=1}^{n} A_i \tag{1}$$

The subset of attributes shared in common between local databases D_i and D_j can be represented as S_{ij} such that

$$S_{ij} = \bigcap_{x=i,j} A_x \tag{2}$$

The union of all attributes shared in common with all local databases forms the shared attribute set S of D.

$$S = \bigcup_{i,j} S_{ij} \tag{3}$$

That means, S contains all the shared attributes within the network.

Our aim is to determine the association rules of global implicit database D with minimum message exchanges. So, the global computation is decomposed into local ones considering the shared attribute constraints as well as preserving the data privacy. Summaries of the local computation will then be aggregated to derive the global association rules. The constraints enforced on sharing attributes among the agents help in ensuring data privacy and confidentiality.
The mathematical formulation of the proposed problem can be described as follows: Let F be a function applied on D and the result be denoted as R, such that,

$$R = F(D). \tag{4}$$

As mentioned earlier, the desired distributed computation here is to derive association rules related to the database D, R denotes the derived associated rule and F denotes the algorithm implementation for deriving R from database D. The responsibilities of an agent is to execute local processing on its database fragment and to communicate with other agents, exchanging local processing results to accomplish the global computation. The shared attribute set S will ultimately result in determining the implicit D obtained from the explicitly involved partitions D_1 to D_n. The functionally equivalent implementation of F (in equation 4) is defined as follows:

$$R(S) = H[h_1(D_1, S), h_2(D_2, S), \ldots, h_n(D_n, S)]. \tag{5}$$

Here, $h_i(D_i, S)$ denotes the local data computation executed by i^{th} agent on database D_i residing on sensor s_i. S denotes the shared attributes and H denotes the aggregation operation upon local computation results, performed by the CH. Each problem requires unique set of h-operators as the count of h_i and the characteristics of both operations H and h_i depend on the shared attribute set S and the involved D_is. The objective of data privacy presented in our method is to prevent attacker/intruder from figuring out any record/tuple at any node in the network and this is made possible by the effective use of hash functions and aggregation methods.

IV. Distributed Mining Association Rules from WSNs

An association rule can be defined as an implication, A implies B, represented as $A \Rightarrow B$, where A, B denotes the item-sets, referred as antecedent and consequent respectively. That means, the transaction records in the database including items in item-set A should also be including items in item-set B. In the proposed technique, determining the association rules in D is the required global computation. It is decomposed and distributed across the network, therefore computation is locally performed and the needed statistical summaries are then collected and exchanged. To initiate the global computation, BS starts the process by requesting the CHs to compute tasks such as support and confidence, then agent at each CH sends request to begin the local computation to its CMs. This will decrease the number and size of messages communicated to and from CMs and their respective CHs and also between BS and CHs, which in turn decrease the consumed energy and prolong the lifespan of the network. Furthermore, this algorithm preserves the data privacy during communication. The proposed distributed algorithms for association rule is composed of three major phases: Initialization, Support and Confidence Computing, and Aggregation. In Initialization phase, every CH creates the relation Shared using shared attributes and shared values from its members. In Support and Confidence Computing phase, CHs initiate the queries and compute the Support and Confidence and send the results to BS and in Aggregation phase, the BS will find the final association rules.

After the WSN is divided into k clusters using any clustering algorithm such as DEC, each cluster head CH_i has m CMs ($s_j^i, j = 1...m$, information of its CMs such IDs, locations, attributes they measure, etc). The CH_i and its members will cooperate with each other and execute mining algorithm as follows:

A. Initialization Phase

We define a relation called P-shared on the attributes in the set S. This relation, P-shared contains tuples related to all combinations of values possible for the attributes in S. The relation P-shared would have mediated the creation of the explicit D, if it was attempted and is used by us in a very similar role. Then, the tuples having zero count at each node will be removed from P-shared and the resulting relation is known as Shared relation.

1) Every CH_i creates *P-shared* relation. The attributes of *P-shared* are the attributes in S and the tuples are the cross product of distinct shared values.
2) Then, generate the *Shared* relation from *P-shared* as follows:
    - Each CM receives s_j^i replies to the $P-Shared$ message by the array $Count-of-Tuples$ that contains count of each tuples.
    - CH_i removes all tuples that have zero count and form *Shared* relation.
    - Index the *Shared* relation beginning with zero

B. Support and Confidence Computing Phase

Support and Confidence Computing phase will be executed by every CH which performs the following two functions:

1) Maintaining the active item-sets and enumerating the candidate sets at the succeeding level from the frequent item-sets in the preceding level.
2) Computing the support and confidence levels.

Implementation of first part can be briefed as follows: the agent on every CH initiates implementation of the algorithm; this agent will carry out the major control tasks of the algorithm, such as finding and managing the active as well as the candidate item-sets, interacting with agents on CMs to compute the two significant measures of association rule namely support and confidence. The computation task is thus decomposed and the process is repeated iteratively and controlled by the agents.

The ratio of the count of transactions containing the item-set to the total transaction count in D is termed as the support of an item-set. Thus, the essential computational primitive in the described technique is to find the total number of tuples in D and it can be calculated only after getting the local computational results from each node. However such computations fulfilling specific attribute-value conditions are a bit complicated and is described as follows.

1) Count of Tuples in Implicit Database: The tuples of global database D are implicit, and are not explicitly visible in a relation, making it more difficult to find the number of enclosed tuples. In our case, we decompose this process of finding tuples, requesting feedback from the agents of D_i's regarding the local counts. The corresponding replies from the agents are then used to determine $N_{total}(D)$, i.e., the total tuples in D. This can be formulated as given below

$$N_{total}(D) = \sum_j \prod_t (N_{D_t})_{cond_j} \tag{6}$$

Here, the subscript $cond_J$ defines the attribute-value condition of the j^{th} tuple of relation Shared, t denotes the index of cluster member and $N(D_t)_{cond_J}$ is the count of tuples in relation D_t satisfying $cond_j$. According to Equation 5, we can have:

$$h_i(D_i, S) = N(D_i)_{cond_j} \tag{7}$$

Such that j refers to the j^{th} tuple of relation Shared. It is required to have such summary for each tuple in Shared from each agent. In order to decrease the interaction between agents, relation Shared can either be managed and maintained by one agent or by each agent separately. The role of the function H is to calculate the sum-of-products from the deduced summaries according to the Equation 6, in which each product term represents the count of tuples satisfying $cond_j$ in a D_i and the resultant gives the number of distinct tuples fulfilling $cond_j$, needed for the implicit Join of all the D_i's. Then the summation operation is performed on the product terms computed for each tuple. This operation simulates a Join operation executed on all the databases without explicit enumeration of the tuples. The most favorable aspect of decomposing $N_{total}(D)$ is that it is possible to translate each product term $N(D_t)_{cond_J}$ into an SQL query; select count (*), such that $cond_j$ can be executed by the local agent at D_t.

2) Support and Confidence for Candidate Sets: The ratio of the count of transactions containing the item-set to the total transaction count in D is termed as the support of an item-set. Also, confidence with respect to a set of transactions refers to the proportion of the transactions that contains A which also contains B. Thus, the essential computational primitive in the described method is to find the total number of tuples in D and it can be calculated only after getting the local computational results from each node. It is possible to extend the decomposition for count into count of tuples satisfying a new condition by changing $cond_j$ in Equation 6 as shown below:

$$N_{new-condition} = \sum_j \prod_{t=1}^{n} N(D_t)_{cond_j \ and \ new-condtion} \tag{8}$$

The above equation is necessary to find the support level for a candidate frequent item-set. The new condition is formed by the attribute values in the frequent item-set. The method in which an agent finds the support measure for a candidate frequent item-set is as described below. In relation Shared, the agent checks the condition specified and identify the tuples matching the attribute-value pairs in the candidate set and then retains those tuples to find the number of tuples resulted from this reduced Shared relation. The support level for a candidate set of attribute-value pairs is given by the ratio of the resultant candidate set count by the total count N_{total}.

The algorithms for generating candidate item-sets and computing the frequent item-sets at every CH, which form the common computational primitives.

C. Extract and Integration of Rules Phase

Extraction: Every CH_i extracts the association rules using frequent item-sets F. The main steps of the extracting rules procedure will be as follows:

- for every frequent item-set $f \in F$, using all nonempty subsets c of f and $c \neq f$.

- for every subset c, if $\frac{support(f)}{support(c)} \geq confidence$
- $R_i = R_i \bigcup \{c \Rightarrow (f - c)\}$

Using Equation 6, calculate the total number of tuples ($no\_of\_tuples = \sum_j \prod_i Count{-}of{-}Tuples[i][j]$) and send message containing the set of rules R_i with confidence of each rule and no_of_tuples to the BS.

Integration: The BS receives the set of rules R_i including the confidence of each rule and the size of the implicit database N_i at CH_i ($i = 1, 2, \ldots, k$). The BS integrates the rules by weighing the confidence of the rules using the N_i and the total number of tuples in the whole database of the network. The main steps of the integration procedure will be as follows:

- Input R_i=$\{r_i^j, c_i^j\}$, $i = 1, 2, \ldots, k$ and j is the number of rules in R_i and c_i^j is the confidence or rule r_i^j.
- N_i is the number of tuples received from CH_i.
- Assume δ is the total number of tuples received from the k clusters.
- $R = R_1 \bigcup \cdots \bigcup R_k$.
- for each $r_l^m \in$ R
- for $i = 1$ to k
- if r_l^m in R_i with confidence c_i^t
- c_l^m= c_l^m + (N_i/δ) * c_i^t

D. Complexity Computing and Analysis

The cost of working with implicitly specified set of tuples can be measured in various ways. One cost model computes the number of messages that must be exchanged among various sites. Complexity for distributed query processing in databases has been discussed in [44] and this cost model measures the total data transferred for answering a query. In our case the amount of data transferred is very little (statistical summaries) but the number of messages exchanged may grow rapidly with the number of iterations of the proposed mining algorithm [45], [46], [48]. At each cluster in the network, in order to extract the association rules, a number of messages need to be exchanged. Let us say:

1) the number of Cluster Heads is K.
2) the average number of members in each cluster is m nodes.
3) we have k -frequent item sets.

We derive below an expression for the number of messages that need to be exchanged for our proposed algorithm dealing with the implicit set of tuples.

Creating the shared relation: the number of messages to create the shared relation can be summarized as follows:

- $m+1$ messages for requesting and receiving the different items of shared attributes.
- $m+1$ messages from CH to members contains Pshared and the reply from members.
- 1 message from CH to members containing the indexed shared relation.

The total number of message per cluster is $3+3m$, i.e., for K clusters it will be $K(3 + 2m)$.

Finding 1^{st} frequent item set: Assume in each cluster, each node has c unshared attributes then we have $c\ m$ unshared attributes. In order to get 1^{st} frequent item, there are two types of items shared and unshared. No messages are needed for shared items where it can find 1^{st} frequent sets using the shared relation at the CH. While for unshared items, CH requests the count of distinct unshared items at each member to get 1^{st} frequent. Therefore, unshared case requires $2c\ m$ messages ($c\ m$ requests to members and $c\ m$ replies from members).

Finding $k-$ frequent item set: As we discussed in algorithm and in example scenario above, each k-frequent item-sets can be in one of the following cases: fully shared, fully unshared and partially shared.

- In case of fully shared, the k-frequent item-sets will be computed at CH, i.e., no exchanged messages are required.
- In case of fully unshared, $2 * k * f_1$ messages are needed to compute the k-frequent item-sets, where f_1 is the average number of frequent item-sets that are fully unshared.
- In case of partially shared, $\sum_{k=2}^{l} 2(k-1)f_2$ exchanged messages are needed to find the k-frequent item-sets in the worst case, where f_2 is the average number of partially shared items and l is the frequent length. In worst case, we have $k - 1$ unshared item-sets, each item-set requires $2(k-1)$ exchanged messages to be sent (requests to members and the replies from members).

  Therefore, the total number of messages will be:

$$Total\ \ messages = K(2m + 3 + 2c\ m + \sum_{k=2}^{l} 2(k-1)f_2 + 2kf_1). \quad (9)$$

V. Simulation Results

In this section, using MATLAB R2016b, we validate feasibility and evaluate the performance of the proposed algorithm. Two types of experiments are performed for validation and evaluated the effectiveness of the proposed approach on the base of DEC [47] as clustering algorithm for WSNs. We test the effect of support value, percentage of CHs and percentage of shared attributes on the number of exchanged messages.

A. Exchanged Messages Parameters

In this set of test 100 sensor nodes are deployed randomly on a 2D-plane to monitor a region with size $100 \times 100\ m^2$. The simulation results are gained by averaging with different topology seeds and 10 clusters (except in the performance evaluations, we vary the number of clusters).

Support value: In the first test, we show the effect of the selected support values on the number of exchanged messages. The support value is varied from 0.1 to 1 with incremental of 0.1 and at each value the number of messages are computed. Fig. 1 shows that the number of messages decreases as the support value increases. The reduction of the number of messages is from 40% to 90% compared to the centralized methodology (or centralized extraction) in which all data transferred to CH, i.e., with zero support value. Therefore, the proposed mining vertical data approach reduces the amount of transferred data and as a result, decreases the number of messages.

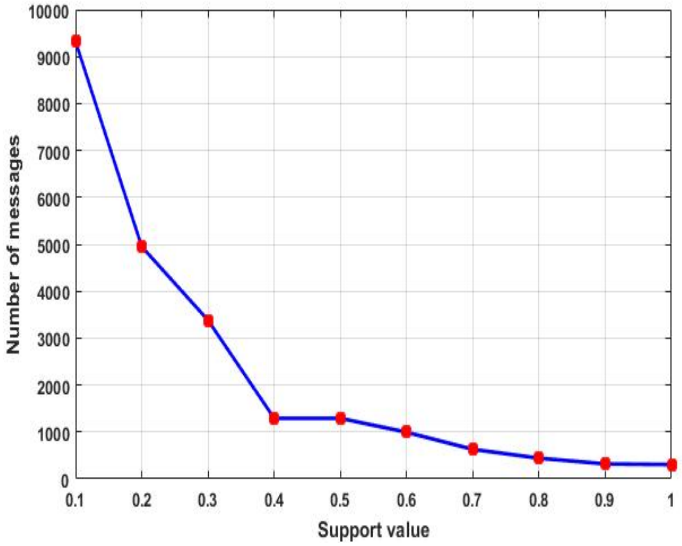


Fig. 1. Number of messages versus support values

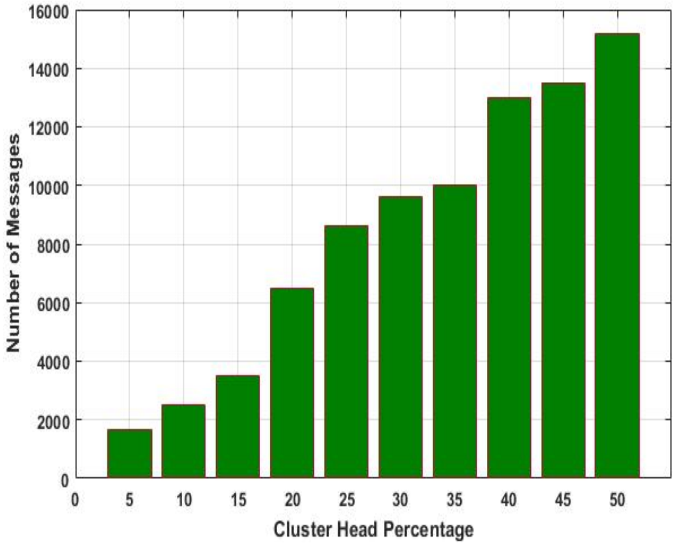


Fig. 2. Percentage of CHs versus number of messages

Cluster head percentage: In the second test, we use the same setting as in previous test except that the percentage of CHs varies from 5% to 50% with incremental of 5% and at each percentage the number of messages are computed. Fig. 2 shows the effect of the CH percentage on the number of messages. It is noted that as the percentage of cluster heads increases the number of messages increases because when the percentage of cluster heads increases, this will increase the messages to base station. Moreover, number of extracted rules increased as number of clusters increased which leads to increase the number of messages.

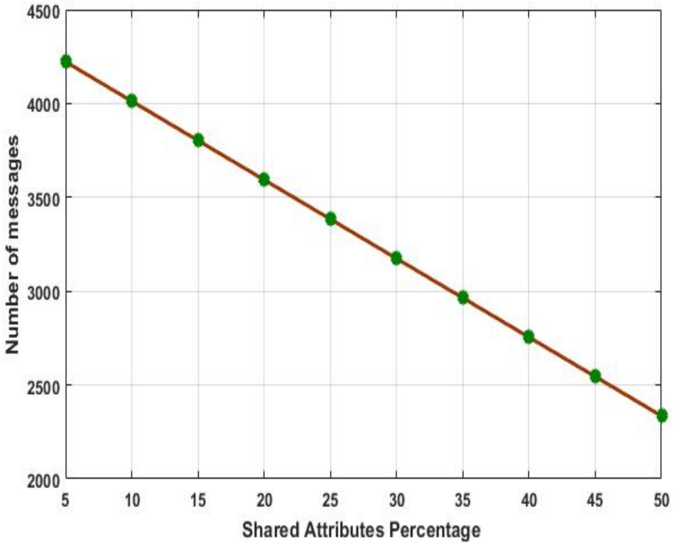


Fig. 3. Number of messages versus percentage of shared attributes

Percentage of shared attributed: In the third test, we use the same setting as in the first test except, the percentage of shared attributes varies from 5% to 50% with incremental of 5% and at each percentage the number of messages are computed. Fig. 3 shows the effect of the number of shared attributes on the number of exchanged messages. We can notice that as the percentage of shared attributes increases, the number of messages decreases because as the number of shared attributes increases, the percentage of unshared attributes decreases and so the number of messages needed for finding and controlling unshared attributes decrease.

VI. Conclusion

In this paper, we have proposed a new approach for mining the sensors data without moving these data to cluster heads or base station to achieve maximum performance in a WSN environment and keep the privacy of the sensor data. The proposed scheme maximizes local computations by utilizing the computing power at each sensor node and reduces the amount of the sensor data transferred in order to decrease the energy dissipation in communication, and thereby the sensor network lifetime is maximized. Moreover, the proposed scheme includes a privacy-preserving technique to ensure the privacy of the sensor data by sending only a summary of the data between the cluster heads and its cluster members and between the cluster heads and the base station. In the future, we will conduct more experiments with different test metrics.

Acknowledgment

The authors gratefully acknowledge the Qassim University, represented by the Deanship of Scientific Research, on the material support for this research under the number (mcs-2018-1-14-S-4006) during the academic year 1440 AH /2019 AD.

References

[1] J. Gama, P. P. Rodrigues, and L. Lopes, "Clustering distributed sensor data streams using local processing and reduced communication," Inteligent Data Analysis, vol. 15, no. 1, pp. 3–28, 2011.

[2] A. Boukerche and S. Samarah, "An efficient data extraction mechanism for mining association rules from wireless sensor networks," in Proceedings of the IEEE International Conference on Communications (ICC '07), pp. 3936–3941, June 2007.

[3] Y. Chi, H. Wang, P. S. Yu, and R. R. Muntz, "Moment: maintaining closed frequent itemsets over a stream sliding window," in Proceedings of the 4th IEEE International Conference on Data Mining (ICDM '04), pp. 59-66,November 2004.

[4] M. Deypir and M. H. Sadreddini, "EclatDS: an efficient sliding window based frequent pattern mining method for data streams," Intelligent Data Analysis, vol. 15, no. 4, pp. 571-587, 2011.

[5] A. Mahmood, K. Shi, and S. Khatoon, "Mining data generated by sensor networks: a survey," Information Technology Journal, vol. 11, pp. 1534-1543, 2012.

[6] J. Rabatel, S. Bringay, and P. Poncelet, "SO MAD: sensor mining for anomaly detection in railway data," in Advances in Data Mining. Applications andTheoretical Aspects, pp. 191-205, 2009.

[7] E. J. Spinosaa, A. P.D. L. F. deCarvalhoa, and J.Gamab, "Novelty detection with application to data streams," Intelligent Data Analysis, vol. 13, no. 3, pp. 405-422, 2009.

[8] P.A. Neves, J.J.P.C. Rodrigues, M. Chen, and A.V. Vasilakos, "A Multi-Channel Architecture for IPv6-Enabled Wireless Sensor and Actuator Networks Featuring PnP Support," J. Netw. Comput. Appl., vol. 37, pp. 12-24, Jan. 2014.

[9] J. M. L. P. Caldeira, J.J.P.C. Rodrigues, and P. Lorenz, "Towards Ubiquitous Mobility Solutions for Body Sensor Networks on HealthCare," IEEE Commun. Mag., vol. 50, no. 5, pp. 108-115, May 2012.

[10] L. Shu, Y. Zhang, G. Min, Y. Wang, and M. Hauswirth, "Cross-Layer Optimization on Data Gathering in Wireless Multimedia Sensor Networks within Expected Network Lifetime J. Univ. Comput. Sci., vol. 16, no. 10, pp. 1343-1367, 2009.

[11] C. Zhu, L. Shu, T. Hara, L. Wang, S. Nishio, and L.T. Yang, "A Survey on Communication and Data Management Issues in Mobile Sensor Networks," Wireless Commun. Mobile Comput., vol. 14, no. 1, pp. 19-36, Jan. 2014.

[12] S.R. Madden, M.J. Franklin, J.M. Hellerstein, and W. Hong, "TinyDB: An Aquisitional Query Processing System for Sensor Networks," ACM Trans. Database Syst., vol. 30, no. 1, pp. 122-173, Mar. 2005.

[13] Y. Yao and J. Gehrke, "Query Processing in Sensor Networks Proc. 1st Biennial Conf. Innovative Data Systems Research (CIDR) Asilomar, Pacific Grove, CA, USA, Jan.5-8, 2003.

[14] M. Umer, L. Kulik, and E. Tanin, "Optimizing Query Processing Using Selectivity-Awareness in Wireless Sensor Networks," J. Comput., Environ. Urban Syst., vol. 33, no. 2, pp. 79-89, 2009.

[15] L. Cheng, Y. Chen, C. Chen, J. Ma, L. Shu, A.V. Vasilakos, and N. Xiong, "Efficient Query-Based Data Collection for Mobile Wireless Monitoring Applications," Comput. J., vol. 53, no. 10, pp. 1643-1657, 2010.

[16] J.G. Pottie and J.W. Kaiser, "Wireless Integrated Network Sensors," Commun. ACM, vol. 43, no. 5, pp. 51-58, May 2000.

[17] R. Kacimi, "Energy Conservation Techniques for Wireless Sensor Networks," Ph.D. dissertation, Dept. Math., Info. and Telecom., INPT Univ., Toulouse, France, Sept. 2009.

[18] C. Dini, "Les Re seaux Capteurs Sans Fil Avec Access Sporadique Au Noeud-puits," Ph.D. dissertation, Info. Eng., Haute Alsace Univ., Mulhouse Cedex, France, Dec. 2010.

[19] E. Fasolo, M. Rossi, J. Widmer, and M. Zorzi, "In-Network Aggregation Techniques for Wireless Sensor Networks: A Survey," IEEE Wireless Commun., vol. 14, no. 2, pp. 70-87, Apr. 2007.

[20] C. Dini and P. Lorenz, "Primitive Operations for Prioritized Data Reduction in Wireless Sensor Network Nodes," in Proc. 4th ICSNC, Porto, Portugal, 2009, pp. 274-280.

[21] C. J. Debono and N. P. Borg, "The Implementation of an Adaptive Data Reduction Technique for Wireless Sensor Networks," in Proc. IEEE ISSPIT, Sarajevo, Bosnia, Dec. 16-19, 2008, pp. 402-406.

[22] L. Shu, J. Lloret, J.J.P.C. Rodrigues, and M. Chen, "Distributed Intelligence and Data Fusion for Sensor Systems," IET Commun., vol. 5, no. 12, pp. 1633-1636, Aug. 2011.

[23] A. Aziz, K. Singh,W. Osamy, and A. M. Khedr, Effective Algorithm for Optimizing Compressive Sensing in IoT and Periodic Monitoring Applications, Journal of Network and Computer Applications, vol. 126, pp. 12-28, 2019

[24] D. M. Omar, "ERPLBC: Energy Efficient Routing Protocol for Load Balanced Clustering in Wireless Sensor Networks, Ad Hoc & Sensor Wireless Networks, vol. 42, pp. 145-169, 2018.

[25] D. M. Omar, and A. M. Khedr, D. P. Agrawal Optimized Clustering Protocol for Balancing Energy in Wireless Sensor Networks, International Journal of Communication Networks and Information Security (IJCNIS) vol. 9, No. 3, pp. 367-375, December 2017.

[26] W. Osamy, A. M. Khedr, An algorithm for enhancing coverage and network lifetime in cluster-based Wireless Sensor Networks, International Journal of Communication Networks and Infor- mation Security (IJCNIS) Vol. 10, No. 1, pp. 1-9, April 2018.

[27] A. M. Khedr, and A. Attia, New Holes and Boundary Detection Algorithm for Heterogeneous Wireless Sensor Networks, International Journal of Communication Networks and Information Security (IJCNIS) vol. 10, No. 1, pp.163-169, April 2018.

[28] A. M. Khedr and D. M. Omar, SEP-CS: Effective Routing Protocol for Heterogeneous Wireless Sensor Networks, Ad Hoc & Sensor Wireless Networks, Vol. 26, pp. 211-232, 2015.

[29] R. Agrawal, R. Srikant, Fast algorithms for mining association rulesProceedings of the 20th International Conference Very Large Data Bases (VLDB '94)1994Citeseer487499

[30] R. J. Bayardo, Efficiently mining long patterns from databasesSIGMOD Record199827285932-s2.0-0032091573

[31] S. Brin, R. Motwani, and C., Silverstein, Beyond market baskets: generalizing association rules to correlationsSIGMOD Record19972622652762-s2.0-0031161999

[32] Cheung, W., Zaiane, O. R.Incremental mining of frequent patterns without candidate generation or support constraintProceedings of 7th International Database Engineering and Applications Symposium2003111116

[33] R. Agrawal, T. Imieliński, and A. Swami, Mining association rules between sets of items in large databasesProceeding of SIGMOD207216

[34] J. Han, J. Pei, Y. Yin, R. Mao, Mining frequent patterns without candidate generation: a frequent-pattern tree approach-Data Mining and Knowledge Discovery20048153872-s2.0-244244995210.1023/B:DAMI.0000005258.31418.83

[35] M. Halatchev, and L. Gruenwald, Estimating missing values in related sensor data streamsProceedings of the 11th International Conference on Management of Data (COMAD '05)2005

[36] N. Jiang, Discovering association rules in data streams based on closed pattern miningProceedings of the SIGMOD Workshop on Innovative Database Research2007.

[37] N. Jiang, L. Gruenwald, Estimating missing data in data streamsAdvances in Databases: Concepts, Systems and Applications20079819872-s2.0-38049151102

[38] K. Loo, I. Tong, B. Kao, Online algorithms for mining inter-stream associations from large sensor networksAdvances in Knowledge Discovery and Data Mining2005291302.

[39] G. S. Manku, R. Motwani, Approximate frequency counts over data streamsProceedings of the 28th International Conference on Very Large Data Bases2002346357

[40] S. K. Chong, S. Krishnaswamy, S. W. Loke, M. Gaber, Using association rules for energy conservation in wireless sensor networksProceedings of the 23rd Annual ACM Symposium on Applied Computing (SAC ’08)March 20089719752-s2.0-5184912666910.1145/1363686.1363911

[41] S. K. Tanbeer, C. F. Ahmed, B. S. Jeong, Y. Lee, Efficient mining of association rules from wireless sensor networksProceedings of the 11th International Conference on Advanced Communication Technology (ICACT ’09)February 20097197242-s2.0-67649882619

[42] A. Boukerche, S. A. Samarah, Novel algorithm for mining association rules in Wireless Ad Hoc Sensor NetworksIEEE Transactions on Parallel and Distributed Systems20081978658772-s2.0-5634909045810.1109/TPDS.2007.70789

[43] K. Romer, Distributed mining of spatio-temporal event patterns in sensor networksProceedings of the 1st Euro-American Workshop on Middleware for Sensor Networks (EAWMS ’06) 2006.

[44] C. Wang, M. Chen, On the Complexity of Distributed Query Optimization. IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 4, pp. 650–662, 1999.

[45] A. M. Khedr, Decomposable Algorithm for Computing k Nearest Neighbors across Partitioned Data, in: International Journal of Parallel, Emergent and Distributed Systems, vol. 31, no. 4, pp. Pages 334-353, 2016.

[46] A. M. Khedr and Raj Bhatnagar, New Algorithm for Clustering Distributed Data using k- means, Computing and Informatics, Vol. 33, pp. 1001-1022, 2014.

[47] F. A. Aderohunmu, J. D. Deng and M. K. Purvis, ”A deterministic energy-efficient clustering protocol for wireless sensor networks,” 2011 Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Adelaide, SA, 2011, pp. 341-346. doi: 10.1109/ISSNIP.2011.6146592

[48] A. M. Khedr, Decomposable Algorithm for Computing k-Nearest Neighbors across Partitioned Data, in: International Journal of Parallel, Emergent and Distributed Systems, vol. 31, no. 4, pp. Pages 334-353, 2016.

[49] S. Dutt, S. Agrawal, and R. Vig. 2018. Cluster-Head Restricted Energy Efficient Protocol (CREEP) for Routing in Heterogeneous Wireless Sensor Networks. Wirel. Pers. Commun. 100, 4 (June 2018), 1477-1497. DOI: https://doi.org/10.1007/s11277-018-5649-x

Efficient Distributed SPARQL Queries on Apache Spark

Saleh Albahli
College of Computer
Qassim University, Saudi Arabia

***Abstract*—RDF is a widely-accepted framework for describing metadata in the web due to its simplicity and universal graph-like data model. Owing to the abundance of RDF data, existing query techniques are rendered unsuitable. To this direction, we adopt the processing power of Apache Spark to load and query a large dataset much more quickly than classical approaches. In this paper, we have designed experiments to evaluate the performance of several queries ranging from single attribute selection to selection, filtering and sorting multiple attributes in the dataset. We further experimented with the performance of queries using distributed SPARQL query on Apache Spark GraphX and studied different stages involved in this pipeline. The execution of distributed SPARQL query on Apache Spark GraphX helped us study its performance and gave insights into which stages of the pipeline can be improved. The query pipeline comprised of Graph loading, Basic Graph Pattern and Result calculating. Our goal is to minimize the time during graph loading stage in order to improve overall performance and cut the costs of data loading.**

***Keywords*—Semantic web; RDF; SPARQL; SPARK; GraphX; triple patterns**

I. Introduction

The semantic web research came a long way from labelling different web pages and linking information to developing better processing systems and efficiently querying semantic web data for information. Today, the semantic web's methods, scale and its representational language are changing drastically. The web data is heterogeneous, it varies in size, semantics and quality of the content which can be true or not. The semantic web data is both high in volume and in velocity due to the wide adoption of digital medium. This has given rise to a lot of research questions. For example, how can we understand data patterns and integrate diverse data for faster access and better quality [1]. The current focus is not to develop new systems from scratch but to improve on existing frameworks with the use of new and improved technologies, such as MapReduce, Apache Spark, big data management, etc.

Resource Description Framework (RDF) is a schema-free data model [2] for linking and storing massive amount of both structured and semi-structured web data in a form of a graph *(subject predicate object)* where *subject* is linked to *object* by *predicate*. With the use of Uniform Resource Identifiers (URIs), RDF links and shares data in a directed and labelled graph where the edges represent the named link and the two nodes represent the two resources or endpoints. This allows data merging even if the underlying schemas are different. Unlike relational or hierarchical data models, RDF stores data in a data graph where there is no concept of roots or hierarchy. It just consists of resources with no single resource having any particular importance over another resource. It perfectly shows the relationship between different resources. Hence, the semantic web is a big global data graph defined in RDF with semantics embeded via RDFS (RDF Schema) [3]. More specifically, semantics triples are added to RDF data using RDFS by applying a set of inference rules to entail new facts, which are not explicitly asserted.

For querying RDF data, SPARQL Protocol and RDF Query Language (SPARQL) is used. SPARQL is currently the standard query language for retrieving and manipulating the data stored in RDF format. SPARQL provide a full range of analytical query operations such as JOIN, SORT and AGGREGATE without requiring a separate schema [4]. There are both programming languages and tools available with which a SPARQL query can be constructed. SPARQL allows to construct queries for RDF data as a set of subject-predicate-object triple. In a relational database terms, it can be considered as a table of three columns i.e. the subject column, the predicate column and the object column. Many RDF data processing systems rely on existing cluster computing engines for parallelized data processing. The current trend of big data needs fast loading and storage strategies to shorten the data ingestion time before query and faster results after querying [5]. Due to this, complex SPARQL queries over large RDF graphs generally have to combine a lot of distributed pieces of data through join operations.

The Apache Jena framework that is widely used as an open source Semantic Web Framework in Java for RDF and provides APIs to extract data from and write to data graph. On the other side, Apache Spark as a MapReduce framework proposes parallel computation using distributed main-memory data abstraction i.e. 1) Resilient Distributed Data Sets (RDD), a distributed lineage supported fault tolerant data abstraction for in memory computations and 2) Data Frames (DF), a compressed and schema-enabled data abstraction [6]. These data abstractions make programming queries easier by enabling translation and processing of high level query expressions such as SPARQL. On top of data abstraction, Spark provide data access models such as GraphX for processing semi-structured data and SPARQL queries over RDF data.

GraphX from Apache Spark is a component for graphs and graph-parallel computation. It reuses Spark's RDD concept, simplifies graph analytics tasks and makes operations on a directed multi-graph. It provides APIs for fast and robust development of a range of algorithms derived from graph theory and applied to search engines and social networks.

Why do we need to improve the query processing of RDF datasets? As mentioned earlier, RDF data has increased in volume and variety available from many different sources on a range of topics. The RDF datasets such as Billion Triples Challenge datasets which are collected by web crawlers, the Linking Open Data Project and the recent conversion of the data.gov dataset are all examples RDF. With the growing availability of huge amount of RDF data, we need efficient ways of querying this data to extract the useful information in a reliable and fast manner [7]. Moreover, the goal is to support different data mining tasks while improving semantic web and exploring vast datasets for innovative insights.

We are achieving this goal by utilizing a cluster's parallelism i.e. our system is able to load and query a large dataset much more quickly than traditional approaches. We have designed experiments to evaluate the performance of several queries ranging from single attribute selection to selection, filtering and sorting multiple attributes in the dataset. We further experimented with the performance of queries using distributed SPARQL query on Apache Spark GraphX and studied different stages involved in this pipeline. We realized that minimizing the data loading time would significantly cut the cost and enhance query performance.

Motivation: RDF is a graph-like representation of knowledge. In this paper, we thus motivate the benefits of Apache Spark's framework GraphX to execute queries on RDF graph data with in memory processing ability of Spark. Our work is adopted the distributed manner and compare it with linear SPARQL query processing with different complexities, readability of queries and scalability using DBpedia dataset.

The importance of this study is that our approach performs better query than traditional approaches even with large datasets. Therefore, our approach of using SPARQL query processing is enhanced with Apache Spark GraphX on different sets of queries using large semantic web datasets. The execution of distributed SPARQL query on Apache Spark GraphX helped us study its performance and gave insights into which stages of the pipeline can be improved. The query pipeline comprised of Graph loading, Basic Graph Pattern and Result calculating. Our goal was to minimize the time during graph loading stage in order to improve overall performance and cut the costs of data loading.

II. Related Study

There are vast amount of research work available in semantic web, rdf, scalable pattern processing and much more. In this section, we will be giving overview of different research papers whose work is relevant to our work.

Chawla et al. [8] processed SPARQL queries on in-memory cluster computing engine Apache Spark and compared SPARQL execution strategies on different query shapes and data sets. They performed experiments on both real-world and synthetic data sets and proposed two new approaches for RDF data model. They were able to achieve a performance improvement by a factor of up to 2.4 on query execution time. The researchers concluded that hybrid query plans combining partitioned and broadcast joins improve query performance. Moreover, using DF representation when RDD exhausted the main memory of the cluster, helped store 10 times more data on the same cluster size with only small loss in performance.

Weaver et al. [6] study the usage of two distributed join algorithms, 1) partitioned join, and 2) broadcast join for the evaluation of basic graph patterns (BGP) using Apache Spark. They suggest through experimentation that hybrid join plans gives more flexibility and achieve better performance than single join plans.

Schätzle [9] aimed to optimize SPARQL queries in order to reduce their execution time. For this purpose, they have modified the conventional All Pair Shortest Path (APSP) algorithms which takes precomputed join costs between triples patterns in a SPARQL query graph using heuristic techniques. Finally, the authors have compared the Floyd Warshall and Johnson algorithms concluding the the former works faster in computing query plans.

Agathangelos et al. [10] devised a novel relational database to efficiently minimize query input size regardless of its pattern shape and diameter. The prototype system called S2RDF is developed on top of Spark and uses relational schema termed as ExtVP (Extended Vertical Partitioning) to execute SPARQL queries. It achieves sub-second runtimes for majority of queries on a billion triples RDF graph.

Naacke et al. [7] developed an application of parallel hash-joins for basic graph pattern matching without the need of any pre-processing, loading or global indexing of the RDF data. The approach relies on the cluster's (using 1024 processors) high bandwidth and fast memory to load and query data in parallel and close to real-time.

Auer et al. [11] discuss existing work in query processing of RDF data using Apache Spark. The RDF data model and the Spark APIs impact the implementation of the RDF quey processing approaches. RDDs have more flexibility for storage and partitioning and GraphX supports graph-parallel and data-parallel data processing.

Table I summarizes the various work conducted in RDF query processing giving details on methodology and findings. In this table, we clearly mention the reference of research work, Apache tool that was used, type of dataset, methodology adopted and finally the findings of each work.

III. Experimental Analysis

We evaluated our proposed model with real world use cases. Following are the details of the evaluation procedure.

A. Experimental Setup

We used two SPARQL implementations to compare and evaluate the performance of semantic queries over RDF datasets. For both of the scenarios we used Google Cloud infrastructure to setup the environment. All of our implementations were written in Java and Scala for performance and extensibility purposes.

1) Linear SPARQL: To establish a baseline for our experiments, we setup Apache Jena on a single node having 7.5 GB of memory, 2 virtual CPUs and 20 GB of persistent storage. Apache Jena is an opensource semantic web framework written in Java and providing APIs to extract data from and write to RDF graphs. Data is first loaded into an abstract model which is then used to query data using SPARQL query language.

TABLE I. COMPARISON TABLE OF PREVIOUS METHODOLOGIES FOR TRIPLE AND GRAPH MODEL.

Reference	Apache Spark Abstraction	Triple/Graph	Methodology	Finding
Naacke et al. (2016) [8]	Apache Spark SPARQL	Triple	Compared five SPARQL query processing strategies over an in-memory based cluster computing engine (Apache Spark) using different query shapes and datasets.	Hybrid query plans combining partitioned join and broadcast joins improve query performance in almost all cases. Moreover, SPARQL Hybrid RDD is bit more efficient than the hybrid DF solution due to the absence of a data compression/decompression overload. We can switch to DF representation if size of the RDDs saturates the main-memory of the cluster.
Naacke et al. (2017) [6]	Apache Spark SPARQL	Graph	Implemented and evaluated four SPARQL query processing strategies over different benchmark queries and data sets	The hybrid query plans combining partitioned and broadcast joins improved query performance in almost all cases and it naturally fits into the recent Spark-based S2RDF system to improve its performance.
Chawla et al. (2017) [9]	Apache Spark SPARQL	Graph	Modified the conventional All Pair Shortest Path (APSP) algorithms which take as input a pre-computed cost matrix of a graph-based SPARQL query.	Optimised SPARQL queries in order to reduce their execution time, APSP's Floyd Warshall algorithm worked faster in computing the plans than the Johnson algorithm..
Schatzle et al. (2016) [10]	Apache Spark SPARQL	Triple / GraphX	A relational partitioning schema for RDF data called ExtVP that uses a semi-join based preprocessing.	It efficiently minimizes query input size regardless of its pattern shape and diameter.
Agathangelos et al. (2018) [11]	Apache Spark, Spark SQL, GraphX, Graph-Frames	Triple / GraphX	Discussion on existing works with efficient query answering and novel ideas for improving query processing by exploiting data parallelization	Data partitioning is a key element for efficient query processing. Graph partitioning focuses on minimizing the edge-cut between partitions. GraphX has not been exploited yet towards this direction and could be an option to build such algorithms.

2) Distributed SPARQL: Distributing SPARQL queries over a cluster of commodity nodes not only improves system performance but will soon become essential as semantic web grows resulting and larger datasets to query and work with.

Apache Spark is one such distributed framework which supports in-memory iterative processing. While vanilla Spark is designed for general purpose distributed processing, it provides a number of libraries which can be used for a wide array of applications. A notable example is GraphX which provides with APIs for working with distributed graphs of nodes and arcs. Having properties pertaining to semantic data processing, GraphX is a perfect candidate to map an RDF graph and perform SPARQL queries in a distributed fashion. We used an open- source implementation S2X; SPARQL query processor for MapReduce based on Spark GraphX.

We used a cluster of 5 nodes to setup Apache Spark in Google Cloud. Each node had 3 GB of memory, 2 virtual CPUs and 20 GB of persistent storage. We used the latest Spark version at the time of writing i.e. Apache Spark version 2.4.3.

3) Dataset: The experiments were designed to ingest RDF datasets in N-Triples format. For datasets which were not in N-Triples format, a preprocessing step was performed to parse and convert the data to N-Triple format. All of the data was uploaded to Google Storage via *gsutil* i.e. google storage API for accessing data on distributed file system.

For smaller dataset, we created a few nodes graph manually. For larger dataset, we used DBPedia dataset [12]. As noted earlier, DBPedia dataset was originally in Turtle format which was preprocessed into N-Triples format before experiments.

B. Experiment Design

We designed several queries to evaluate the performance on available datasets. These ranged from simplest query on single attribute selection to selection, filtering and sorting on multiple attributes in the dataset.

The simplest query was to select person names in the dataset.

```
// Simple SPARQL query
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name
WHERE {?person foaf:name ?name .}
```

The second and comparatively complex query was on DBPedia dataset.

```
// Complex SPARQL query
SELECT *
WHERE {
?person
    <http://dbpedia.org/ontology/deathDate>
?deathDate .
?person <http://xmlns.com/foaf/0.1/page>
    ?page .
FILTER(?deathDate >= "1941-01-01"^^xsd:date)
FILTER(?deathDate <= "1942-01-01"^^xsd:date)
} order by ?deathDate;
```

Another query on DBPedia dataset was:

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX res: <http://dbpedia.org/resource/>
PREFIX rdf:
<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:
<http://www.w3.org/2000/01/rdf-schema#>
SELECT COUNT(DISTINCT ?uri)
WHERE {
?uri rdf:type dbo:Film .
?uri dbo:starring res:Leonardo_DiCaprio .
}
```

Next we studied the execution of distributed SPARQL query on Apache Spark GraphX in terms of performance. This gave us insights into which stages can be improved further explained in Section III-D

C. Distributed SPARQL Performance

Fig. 1 shows the performance comparison results for linear SPARQL query processing and its distributed variant. As we can see, for smaller datasets, Linear query processing outperforms distributed processing. This happens mainly because large scale distributed systems do have overheads associated with system initialization and communication costs which only gets mitigated by processing enough data. This also enforces the fact that semantic processing needs to embrace parallel and distributed processing principals as the data to process keeps on increasing.

During the experiments, we did note that as the dataset increased, Apache Jena kept getting out of memory. We were able to handle the situation by increasing JVM's heap size. However, this is a temporary solution and there is a limit to size of heap on a single node. This again shows that single node processing for SPARQL queries on semantic data is by no means optimal.

D. Distributed SPARQL Analysis

To further study Spark based distributed SPARQL query processing, we analyzed different stages involved in the distributed SPARQL pipeline. Fig. 2 shows the percentage time spent in different stages while executing SPARQL query in a distributed environment using Apache Spark's GraphX library. In total, the query pipeline consists of three separate stages:

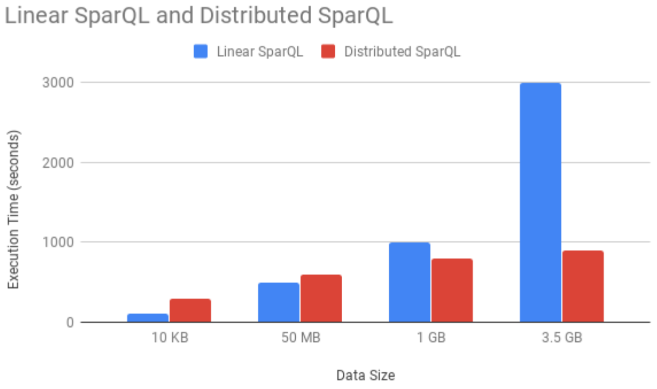


Fig. 1. Performance comparison for Linear vs. Distributed SPARQL processing with increasing data size

(1) Graph loading, (2) Basic Graph Pattern, and (3) Result calculating. The most expensive stage in terms of performance is Graph loading stage. This is expected since the first stage involves reading data from distributed storage into the memory along with all the graph creation logic. As illustrated in Fig. 2, the first stage takes almost 65% of the total execution time. Once data is in memory, the next step is to create the graph logic including the creation of *Triple Patterns*. Finally, the last step is the actual SPARQL query processing on the dataset and getting the result of the query.

These stages also present a potential improvement over the current execution pipeline. Since most of the time is spent in IO during the first stage, the system can be kept alive to keep data in memory and avoid reloading data for every query. This would greatly cut the cost of data loading, enhancing the over all query performance.

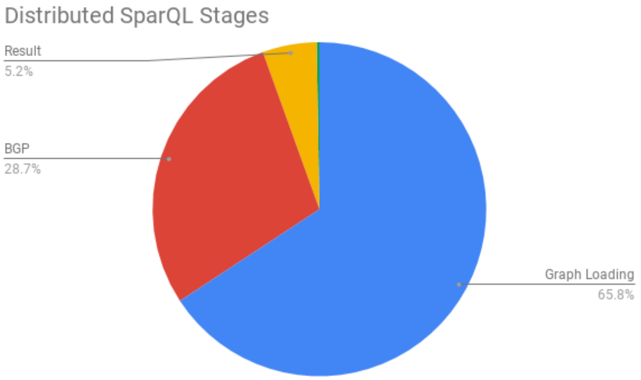


Fig. 2. Analysis of different stages for distributed SPARQL

E. Data Loading

Time spent in IO is a limiting factor for most in- memory systems. To study this, we compared the IO time for both linear and distributed implementation. Fig. 3 shows the results of this experiment. With default JVM settings, linear SPARQL failed to handle larger datasets since it resulted in out of memory exceptions. However, this did not happen in distributed setup since data was partitioned among memory in distributed nodes.

For all practical purposes, user can always tweak the JVM heap size to avoid such out of memory exceptions. However, as discussed earlier as well, although it is doable, it is not a desirable solution since it involves manual configuration on the user's end.

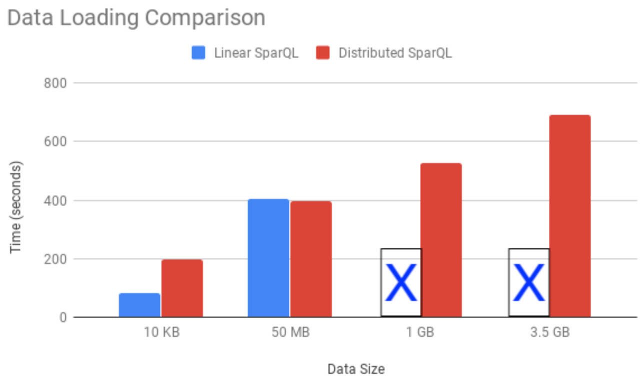


Fig. 3. Data Loading Time Comparison with default JVM settings

F. *Query Complexity*

Query complexity can be defined as the particular function a query performs. Although, an infinite set of queries can be designed for different datasets, we categorize them into three representative queries. The simplest query simply projected a few columns from the desired dataset. The second used an aggregation function on the dataset to compute frequency, average etc. of a particular column. The last query involved joins on different dataset columns.

In our final experiment, we compared the query performance with different complexities having fixed dataset. Fig. 4 shows the results for different sets of queries for a fixed dataset size of 1 GB. As expected, the results for simple projection and aggregation are much better for distributed SPARQL mainly because, backed distributed framework is able to work on partitioned data in parallel on distributed nodes. A linear query execution on the other hand has limited parallelization and can limited performance. However, for queries involving joins, distributed environments can limit the parallelization since records from multiple dataset partitions may potentially need to be co-grouped and joined. This involves extra IO and serialization overheads associated with moving data across network. Therefore, as shown in the Fig. 4, the performance gain for join queries is a limited as compared to other queries.

IV. Conclusion and Future Works

The obtained results showed a good query response time while using Spark based SPARQL comparing with Jena baseline performance results. We recognized that reducing the time of loading data will lead to lower the cost of potentially expensive query processing and hence improve query performance. Moreover, distributed SPARQL queries achieve better response time handling larger datasets than when running on linear SPARQL as data was partitioned among memory in distributed nodes. This is unlike the linear SPARQL which trigger out of memory exceptions. However, we can always tweak the JVM heap size to avoid such out of memory exceptions but it is still not a suitable solution as it involves manual configuration on the user's end. Since our work focus on two SPARQL queries: linear and distributed, it would be interesting in future to extend the study to investigate different types of queries as well. Furthermore, integrating parallel architecture with new settings and hardwares using different parallel RDF queries can be further investigated to bring new outstanding results in the field of semantic web.

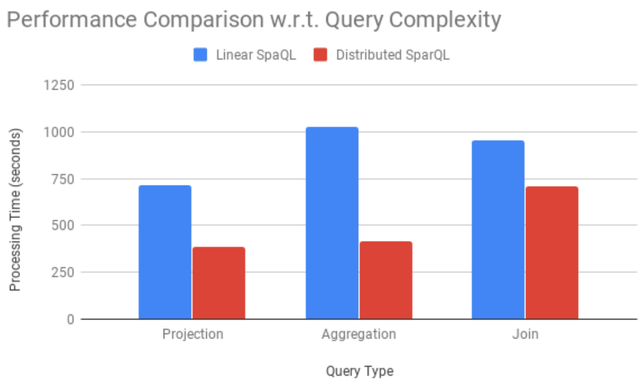


Fig. 4. Query performance comparison for different queries.

References

[1] A. Bernstein, J. Hendler, and N. Noy, "A new look at the semantic web," *Commun. ACM*, vol. 59, no. 9, pp. 35–37, Aug. 2016. [Online]. Available: http://doi.acm.org/10.1145/2890489

[2] F. Manola, E. Miller, B. McBride *et al.*, "Rdf primer," *W3C recommendation*, vol. 10, no. 1-107, p. 6, 2004.

[3] D. Brickley, R. V. Guha, and B. McBride, "Rdf schema 1.1," *W3C recommendation*, vol. 25, pp. 2004–2014, 2014.

[4] S. Albahli and A. Melton, "Triplefca: Fca-based approach to enhance semantic web data management," in *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1. IEEE, 2016, pp. 625–630.

[5] T. Chawla, G. Singh, and E. S. Pilli, "Hypso: Hybrid partitioning for big rdf storage and query processing," in *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. ACM, 2019, pp. 188–194.

[6] H. Naacke, B. Amann, and O. Curé, "Sparql graph pattern processing with apache spark," in *Proceedings of the Fifth International Workshop on Graph Data-management Experiences & Systems*. ACM, 2017, p. 1.

[7] J. Weaver and G. T. Williams, "Scalable rdf query processing on clusters and supercomputers," in *The 5th international workshop on scalable semantic web knowledge base systems (ssws2009)*, vol. 8, 2009.

[8] H. Naacke, O. Curé, and B. Amann, "Sparql query processing with apache spark," *arXiv preprint arXiv:1604.08903*, 2016.

[9] T. Chawla, G. Singh, and E. S. Pilli, "A shortest path approach to sparql chain query optimisation," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017, pp. 1778–1778.

[10] A. Schätzle, M. Przyjaciel-Zablocki, S. Skilevic, and G. Lausen, "S2rdf: Rdf querying with sparql on spark," *Proceedings of the VLDB Endowment*, vol. 9, no. 10, pp. 804–815, 2016.

[11] G. Agathangelos, G. Troullinou, H. Kondylakis, K. Stefanidis, and D. Plexousakis, "Rdf query answering using apache spark: Review and assessment," in *2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW)*. IEEE, 2018, pp. 54–59.

[12] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.

A Modular Aspect-Oriented Programming Approach of Join Point Interfaces

Cristian Vidal[1*], Erika Madariaga[2*], Claudia Jiménez[3*], and Luis Carter[4*]

[1]Departamento de Administración, Facultad de Economía y Administración, Universidad Católica del Norte, Antofagasta, Chile,
[2]Ingeniería Informática, Facultad de Ingeniería, Ciencia y Tecnología, Universidad Bernardo O'Higgins, Santiago, Chile,
[3]Ingeniería Civil Informática, Escuela de Ingeniería, Universidad Viña del Mar, Viña del Mar, Chile,
[4]Ingeniería Civil Industrial, Facultad de Ingeniería, Universidad Autónoma de Chile, Chile,

***Abstract*—This paper describes and analyzes the main differences and advantages of the Join Point Interfaces (JPI) as an Aspect-Oriented Programming (AOP) approach for the modular software production concerning the standard aspect-oriented programming methodology for Java (AspectJ) to propose a structural modeling approach looking for modular software solutions. Using a Software Engineering point-of-view, we highlight the relevance of structural and conceptual design for JPI software applications. We model and implement a classic example of AOP using AspectJ and JPI as an application example to review their main difference and highlight the JPI consistency between products (models and code). Our proposal of UML JPI class diagrams allows the definition of oblivious classes which know about their JPI connections, an essential element to adapt and transform tradition like-AspectJ AOP solutions to their JPI version. Thus, for the modular software production and education, JPI seems an ideal software development approach.**

***Keywords*—Aspect-Oriented Programming; AspectJ; JPI; class diagrams; UML**

I. Introduction

The methodology of Aspect-Oriented Programming (AOP) [1] allows to encapsulate or modularize the so-called “non-modularizable cross-concerns” by classical programming methodologies such as object-oriented programming and structured programming. The cross-functionalities are functions spread as part of the modules such as in the methods of classes mixing their steps and nature. In this way, the cross-functionalities represent non-modularizable functional elements by traditional software development methodologies like structured and object-oriented software development. We talk of aspect-oriented software development given the adaptation of other phases in the software development process to encapsulate the so-called cross-concerns.

The users' authentication and registration or log of their actions are classic examples of cross-concerns [2] [3] [4]. By modularizing cross-concerns of an object-oriented software system, classes and their methods can respect the single-responsibility principle [5]), that is, classes and their methods own and know their function and well-defined purpose. Thus, consistency should exist in the behavior of classes' methods which do not presents actions out of their primary objective. In summary, the principle of single-responsibility states that a class should encapsulate only one responsibility. This represents a fundamental principle of modularization and object-oriented programming [6].

Fig. 1 shows a diagram of UML use cases that represents a system and two cases of use that are examples of cross-concerns, usually present in software systems: use cases ‘Logging’ and ‘Authentication’. Note that Jacobson [7] [8], and Vidal et al. [2] indicate the use of ‘extends’ associations between use cases that represent cross-concerns and base-use cases. In traditional software systems, it is common that before acting, the user must authenticate, and if such authentication does not occur, the action does not proceed. Besides, it is usual for current information systems to keep a record of the actions performed by their users. In this context, Fig. 2 presents a traditional UML class diagram for the example system of Fig. 1.

Fig. 2 shows two classes, Class1 and Class2, with a one-to-many association. Each class has two attributes and four methods. We can appreciate that both classes present the methods A (..) and B (..), that is, Register Actions and Authenticate. Assuming that the functionalities and associated behavior of the methods A(..) and B(..) are not specific to Class1 and Class2, and how these functionalities cannot be represented in independent classes so that Class1 and Class2 respect the principle of only responsibility, then methods A(..) and B(..) are examples of cross-concerns.

Table I presents the Java code associated with the UML class diagram of Fig. 2. As can be seen in both classes, public methods explicitly invoke the execution of methods A(..) and B(..) which does not correspond to the responsibility nature of public methods of Class1 and Class2, respectively. That is, neither the classes nor the methods of this example respect the principle of sole responsibility.

Such as Kiczales et al. [1] pointed out, classic or traditional AOP permits the elimination of cross-concerns in classes of object-oriented software systems modularizing them as aspects. However, as Bodden [9] points out, Instroza et al. [10], and Bodden et al. [11], traditional AOP Aspect-style does not permit achieving a complete modularization for the existence of implicit dependencies between aspects and classes. For getting software products with greater modularity, the works of [9] [10] [11] present Join Point Interfaces (JPI) to eliminate implicit dependencies between classes and aspects of classic AOP AspectJ-style.

Considering the mentioned JPI benefits for developing

*Corresponding author

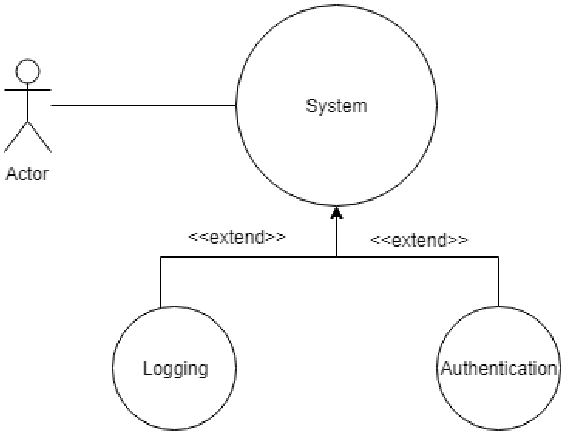


Fig. 1. Example of Cross-Concern for Logging the Actions and Authenticating in a Use Cases Diagram of a System.

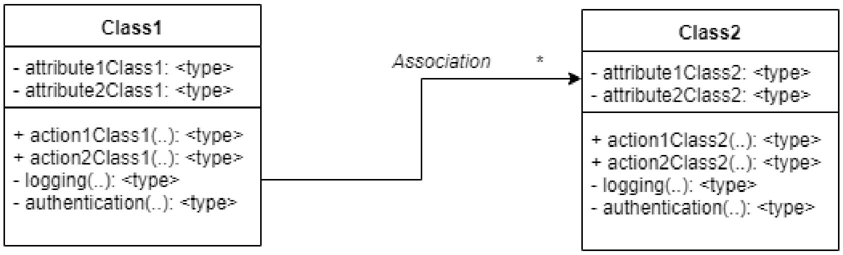

Class1
- attribute1Class1: <type>
- attribute2Class1: <type>
+ action1Class1(..): <type>
+ action2Class1(..): <type>
- logging(..): <type>
- authentication(..): <type>
Association
*
Class2
- attribute1Class2: <type>
- attribute2Class2: <type>
+ action1Class2(..): <type>
+ action2Class2(..): <type>
- logging(..): <type>
- authentication(..): <type>

Fig. 2. Example of UML Class Diagram with Cross-Concerns.

TABLE I. JAVA EXAMPLE WITH CROSS-CONCERNS.

public class Class1 {
public <type>attribute1Class1;
public <type>attribute2Class1;
 	
public <type>action1Class1([Arg1,., ArgN]){
authentication(..);
...
logging(..);
}
 	
public <type>action2Class1([Arg1,., ArgN]){
authentication(..) ;
...
logging(..);
}
}
	public class Class2 {
	public <type>attribute1Class2;
	public <type>attribute2Class2;
	
	public <type>action1Class2([Arg1,., ArgN]){
	authentication(..);
	...
	logging(..);
	}
	
	public <type>action2Class2([Arg1,., ArgN]){
	authentication(..);
	...
	logging(..);
	}
	}

modular software, the main objective of this article is to apply JPI concepts on a base example to demonstrate and describe its practical advantages for getting modular aspect-oriented software solutions without implicit dependencies. Moreover, this article proposes an approach for the structural modeling of JPI solutions along with gives the bases for future research for behavioral modeling of JPI solutions.

This work organizes as follows: next section summarizes related works. Section 3 presents and exemplifies the AspectJ and JPI aspect-oriented programming approaches. Section 4 describes and exemplifies UML class diagrams along with to propose and exemplify a UML class diagram for JPI solutions. Finally, conclusions and future research work regarding behavioral modeling ideas using UML sequence diagrams for JPI solutions.

II. RELATED WORKS

Different works about AOP applications and extensions already exist such us [12] and [13] which define formal rules to specify AOP solutions and [14] that uses an AOP framework to monitor the run-time state and behavior of software applications. The works of [9] [10] and [11] are bases and programming background of the JPI framework for building AOP solutions. The work of [15] represent one of the first article to describe and exemplify JPI modeling ideas. This article describe an exemplify a more precised and detailed structural modeling approach for JPI solutions which represent advances to look for a complete JPI software development process.

III. Aspect-Oriented Programming

This section summarizes AspectJ and JPI AOP styles.

A. AspectJ-Style AOP

Kiczales et al. [1] argue that the aspects, advice units, introductions, relations between data-types, join points, and pointcuts represent relevant characteristics of traditional AspectJ-like AOP. Also, there is a difference between base elements and aspects, which are usually associated. Pointcuts in aspects define the relations between aspects and classes, and aspects can advise methods related to their pointcuts, or through the inclusion of new behavior and attributes in advised classes, that is, by introductions or declarations among types in those classes.

In classical AOP, base elements of a system are oblivious about their interaction with aspects of the system, as well as the possibility of being advised, that is, base elements know nothing about any change or new behavior. In this way, a class that does not expect interruptions and changes in its functionality may experience unexpected changes. According to Bodden et al. [11], this is one of the implicit dependencies that exist in classic AOP as well as in any previous AOP like AspectJ modeling proposal.

Implicit dependencies between software modules complicate the software development process. Eliminating implicit dependencies between aspects and classes would facilitate the software development by independent groups, one in charge of system classes and another one in charge of aspects of the system. That approach seems adequate for software evolution. The work of [11] indicate that a solution is a complete knowledge of the classes of the system by the developers of aspects, and also, constant communication between development teams is necessary before any change in the base elements, as well as, to indicate all already advised class elements. This complete communication seems viable in small software development teams, but it is not always possible to develop software between large and multiple independent teams.

AspectJ uses execution(..) in the definition of pointcuts [16] [8] to capture defined objects which execute one of its methods. For example, by using execution(type C2.met (..)) && this (obj1) && target (obj2), obj1 and obj2 represent the same object of class C2 that executes the method met (..). In the same way, as pointed out by [16] and [8], in addition to capturing the object that executes a method, it is also possible to capture the object that invokes the execution of a method of a given object, which could be the same. For example, in the definition of the cutoff point call (type C2.met (..)) && this (obj1) && target (c2), obj1 represents the object that invokes the call of the method met (..) of an object of class C2, while obj2 represents the object of class C2 that executes the method met (..).

Table II presents the AOP AspectJ code solution for the example code of Fig. 3. As shown in this figure, the principle of sole responsibility is respected in Class1 and Class2 classes. In this example, the methods, methods action1Class1(..) and action2Class1(..), as well as action1Class2(..) and action2Class2(..), of the classes Class1 and Class2 respectively, are oblivious concerning the inclusion of aspects behavior. This ingenuity of the advised methods is problematic for methods that must preserve intact their original behavior. That represents one of the implicit dependencies between classes and previously mentioned aspects. In the same way, if the signature of one of the public methods of Class1 or Class2 changed, potentially Aspect1 and Aspect2 would not be effective, that is, there would be no join point, and besides, no compilation errors exist. The following subsection describes the JPI approach proposed by [10] and [11] to look for eliminating these dependencies between aspects and classes.

B. JPI

Such as [15] remark, the main difference of JPI concerning the classic POA is, as its name indicates, the use of join point interfaces as intermediate points of the association between classes and aspects. In this way, JPI allows the elimination of oblivious classes since advisable classes explicitly exhibit join point interfaces and define pointcut rules for the effectiveness of those unions. In the same way, aspects implement those interfaces and do not know directly about the advisable classes. Therefore, JPI eliminates the implicit dependencies between classes and aspects which permit reaching higher levels of modularization regarding classical AOP. JPI, as an extension of AspectJ, supports traditional AspectJ code to facilitate the adaptation of AspectJ code to JPI. JPI, like traditional AOP, allows declaration between types of classes in aspects which do not require explicit join point interfaces, that is, classes continue being oblivious regarding the introduction of attributes and behavior by aspects.

Table III shows a JPI solution for the classes and aspects of the AOP AspectJ-style example of Table II. Table III presents a new code box for the join point interfaces JPIAauthentication and JPILogging regarding the classes and aspects of Table II. As seen in Table III, each aspect, to be effective advising classes, requires implementing join point interfaces exhibited by those classes. Thus, Aspect1 implements the JPIAuthentication and Aspect2 implements the JPILogging join point interfaces respectively, both exhibited by Class1 and Class2.

As mentioned above, a join point interface can be used for the inclusion of new methods and attributes, that is, for the inter-type declaration in oblivious classes. JPI also allows defining global join point interfaces for an AspectJ style of AOP (Bodden et al. 2014).

We can appreciate in the JPI code of Table III, classes Class1 and Class2 explicitly exhibit the join point interfaces for the execution of any of their public methods. In this way, JPI allows the elimination of implicit dependencies between classes and aspects of traditional AOP. First, non-oblivious classes indicate their methods associated with a join point interface, that is, their advisable methods. Second, if a class that exhibits join point interfaces regarding some method's signature and that method undergoes some signature change

TABLE II. AOP AspectJ Code for the Example.

<table>
<tr>
<td>public class Class1 {
public <type>attribute1Class1;
public <type>attribute2Class1;

public <type>action1Class1([Arg1,., ArgN]){
...
}

public <type>action2Class1([Arg1,., ArgN]){
...
}
}</td>
<td>public class Class2 {
public <type>attribute1Class2;
public <type>attribute2Class2;

public <type>action1Class2([Arg1,., ArgN]){
...
}

public <type>action2Class2([Arg1,., ArgN]){
...
}
}</td>
</tr>
<tr>
<td>public aspect Aspect1 {
pointcut pcAuthentication(..):
execution(<type>Class1.action1Class1(
[Arg1, .., ArgN])) ||
execution(<type>Class1.action2Class1(
[Arg1, .., ArgN])) ||
execution(<type>Class2.action1Class2(
[Arg1, .., ArgN])) ||
execution(<typo>Class2.action2Class2(
[Arg1, .., ArgN])) ... ;

before(..): pcAuthenticate(..){
... //Authentification
}
}</td>
<td>public aspect Aspect2 {
pointcut pcLogging(..):
execution(<type>Class1.action1Class1(
[Arg1, .., ArgN])) ||
execution(<type>Class1.action2Class1(
[Arg1, .., ArgN])) ||
execution(<type>Class2.action1Class2(
[Arg1, .., ArgN])) ||
execution(<typo>Class2.action2Class2(
[Arg1, .., ArgN])) ... ;

after(..): pcLoggin(..){
... //Logging
}
}</td>
</tr>
</table>

without updating the pointcut rule in the JPI exhibition, then a compilation error occurs. That is, the class developer team must indicate changes in the methods' signature in the associated joint point interface exhibition. Then, in JPI, with a clear definition of join point interfaces, classes and aspects development teams could exist.

IV. JPI UML Class Diagram: Proposal and Application

A UML class diagram represents the classes of a software system along with their associations [17]. Such as Vidal et al. 2015 [15] and Torres et al. [18] argue, a UML class diagram allows classifying classes and their associations through the use of stereotypes besides. For example, usually, a class interface is represented by an <<interface>> stereotyped class. In this way, UML class diagrams seem suitable for the representation of JPI solutions.

Next, we define rules and names of elements for a JPI UML class diagram:

- Classes and their associations are defined in the usual way as in a UML class diagram.
- A join point interface is declared with the stereotype <<jpi>> or <<global jpi>> depending on whether the advisable classes explicitly or implicitly exhibit those interface, respectively. In this proposal of JPI UML class diagrams, a join point interface does not have attributes either methods, that is, a JPI interface represents a method without a signature.
- An aspect, which is a stereotyped class with <<aspect>>, allows to define a series of variables and methods of aspects, as well as to define methods of interfaces of point of union, and declarations between types or introductions.
- Classes can exhibit interfaces of stereotyped junctions with <<jpi>>. In this way, when a class exhibits a junction point interface, there is an association of the class to a join point interface. The role of the class of this association presents a first line with the stereotype <<exhibits>> together with the signature of the interface, and a second line with the pointcut rule.
- A global join point interface includes an association to the recommended class with a line that indicates the signature of the join point and another line with the definition of the pointcut rule.
- The aspects, to effectively advise this is, to add behavior on the call or execution of methods of advisable classes, they must implement join point interfaces. For this reason, each aspect, to be effective, presents an association towards the associated point of attachment interfaces. The role of the aspect in these associations is 'implements'.
- The aspects allow inter-type declaration, that is, to add attributes and methods to existing classes. Thus, an association between aspects and classes is used, where the role of the aspect is 'adding'.

TABLE III. AOP JPI CODE FOR THE EXAMPLE.

```
public class Class1 {
    exhibits JPIAuthentication(..):
        execution(* action1Class1([Arg1, ArgN]) ||
        execution(* action2Class1([Arg1, ArgN])) &&
        args(..);

    exhibits JPILogging(..):
        execution(* action1Class1([Arg1, ArgN]) ||
        execution(* action2Class1([Arg1, ArgN])) &&
        args(..);

    public <type>attribute1Class1;
    public <type>attribute2Class1;

    public <type>action1Class1([Arg1,., ArgN]){
        ...
    }

    public <type>action2Class1([Arg1,., ArgN]){
        ...
    }
}
```

```
public class Class2 {
    exhibits JPIAuthentication(..):
        execution(* action1Class2([Arg1, ArgN]) ||
        execution(* action2Class2([Arg1, ArgN])) &&
        args(..);

    exhibits JPILogging(..):
        execution(* action1Class2([Arg1, ArgN]) ||
        execution(* action2Class2([Arg1, ArgN])) &&
        args(..);

    public <type>attribute1Class2;
    public <type>attribute2Class2;

    public <type>action1Class2([Arg1,., ArgN]){
        ...
    }

    public <type>action2Class2([Arg1,., ArgN]){
        ...
    }
}
```

```
jpi JPIAuthentication(..);
jpi JPILogging(..);
```

```
public aspect Aspect1 {
    before JPIAuthentication(..){
        ... //Authentication
    }
}
```

```
public aspect Aspect2 {
    after JPILogging(..){
        ... //Logging Actions
    }
}
```

Because the proposed JPI UML class diagram extension considers the use of stereotypes and special keywords, any UML design tool can be used for the JPI UML diagram design. In this way, it is possible to model a JPI solution and system structurally. Fig. 3 presents an application of this proposal of UML JPI class diagrams on the JPI example of Table III. Note that JPIInterfaceA corresponds to JPIAauthentication while JPIInterfaceB corresponds to JPILogging. Similarly, AspectA corresponds to Aspect1 and AspectB with aspect2 of Table III. As Fig. 3 shows, there is a clear analogy between the number of components in the JPI UML class diagram and the JPI code solution. Precisely, to review the consistency and modular advantages of this proposal of UML JPI class diagrams is part of the future work for the authors of this work.

It should be noted that for traditional AOP UML class diagram, there are already proposals such as Kojarski et al. [19] and [20] which use existing UML modeling tools.

V. CONCLUSIONS

JPI makes possible the generation of aspect-oriented solutions without implicit dependencies, which in turn allows achieving a high degree of modularization concerning traditional AOP. As this paper mentioned, JPI enables the definition of introductions without join point interfaces, that is, for oblivious classes of the introduction of new attributes and behavior, however, those classes are no longer oblivious about changes in the behavior of their methods through aspects' advice units.

For the structural modeling of JPI applications, this work extends UML class diagrams using JPI concepts for the modeling JPI solutions. As presented in the modeling example, our UML class diagrams proposal for JPI captures basic elements of JPI such as global or non-global join point interfaces. Other elements of JPI, such as closure and generic join points [9], [10], [11], are part of future extensions to this modeling proposal. Also, this proposal of UML JPI class diagrams allows defining oblivious classes, which is an essential element to achieve a complete adaptation and transformation of AOP solutions into JPI solutions.

As future work, the authors of this article work on a complete proposal of structural modeling for JPI applications, as well as on ideas for modeling the behavior of JPI systems through UML sequence diagrams. For this last idea of future work, the actors of each scenario are identified, where the aspects are clear participants and, for which, the participating objects communicate in the existence of join points, that is,

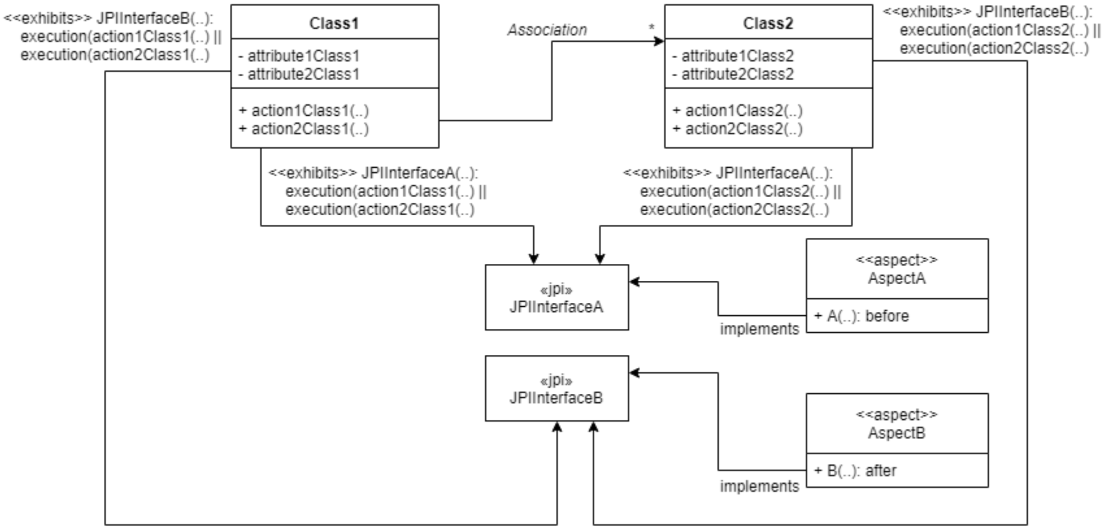

<<exhibits>> JPIInterfaceB(..):
execution(action1Class1(..) ||
execution(action2Class1(..)
Class1
- attribute1Class1
- attribute2Class1
+ action1Class1(..)
+ action2Class1(..)
Association
*
Class2
- attribute1Class2
- attribute2Class2
+ action1Class2(..)
+ action2Class2(..)
<<exhibits>> JPIInterfaceB(..):
execution(action1Class2(..) ||
execution(action2Class2(..)
<<exhibits>> JPIInterfaceA(..):
execution(action1Class1(..) ||
execution(action2Class1(..)
<<exhibits>> JPIInterfaceA(..):
execution(action1Class2(..) ||
execution(action2Class2(..)
«jpi»
JPIInterfaceA
<<aspect>>
AspectA
+ A(..): before
implements
«jpi»
JPIInterfaceB
<<aspect>>
AspectB
+ B(..): after
implements

Fig. 3. Example of JPI UML Class Diagram.

respecting the pointcut rules. The idea is to model, through UML diagrams, the structure, and behavior of JPI solutions, to work on the generation of JPI code from the models. From a Software Engineering point-of-view, we want to use JPI as an approach for the modular and consistent software development approach.

REFERENCES

[1] G. Kiczales, "Aspect-oriented programming," *ACM Comput. Surv.*, vol. 28, no. 4es, Dec. 1996.

[2] C. V. Silva, R. Saens, C. D. Río, and R. Villarroel, "Aspect-oriented modeling: Applying aspect-oriented UML use cases and extending aspect-z," *Computing and Informatics*, vol. 32, no. 3, pp. 573–593, 2013.

[3] C. V. Silva, R. Villarroel, and C. P. Vasquez, "Jpiaspectz: A formal specification language for aspect-oriented JPI applications," in *33rd International Conference of the Chilean Computer Science Society, SCCC 2014, Talca, Maule, Chile, November 8-14, 2014*, 2014, pp. 128–131.

[4] C. V. Silva, R. Villarroel, R. S. Simón, R. Saens, T. Tigero, and C. D. Río, "Aspect-oriented formal modeling: (aspectz + object-z) = ooaspectz," *Computing and Informatics*, vol. 34, no. 5, pp. 996–1016, 2015.

[5] D. Wampler, *Aspect-oriented design principles: Lessons from object-oriented design*, 1st ed. Vancouver, Canada: Proceedings of the Sixth International Conference on Aspect-Oriented Software Development (AOSD'07), 2007.

[6] R. C. Martin, *Agile Software Development: Principles, Patterns, and Practices*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2003.

[7] I. Jacobson, "Use cases and aspects-working seamlessly together," *Journal of Object Technology*, vol. 2, no. 4, pp. 7–28, 2003.

[8] I. Jacobson and P.-W. Ng, *Aspect-Oriented Software Development with Use Cases (Addison-Wesley Object Technology Series)*. Addison-Wesley Professional, 2004.

[9] E. Bodden, "Closure joinpoints: Block joinpoints without surprises," in *Proceedings of the Tenth International Conference on Aspect-oriented Software Development*, ser. AOSD '11. New York, NY, USA: ACM, 2011, pp. 117–128.

[10] M. Inostroza, E. Tanter, and E. Bodden, "Join point interfaces for modular reasoning in aspect-oriented programs," in *Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering*, ser. ESEC/FSE '11. New York, NY, USA: ACM, 2011, pp. 508–511.

[11] E. Bodden, E. Tanter, and M. Inostroza, "Join point interfaces for safe and flexible decoupling of aspects," *ACM Trans. Softw. Eng. Methodol.*, vol. 23, no. 1, pp. 7:1–7:41, Feb. 2014.

[12] C. Vidal Silva, R. Saens, C. Del Río, and R. Villarroel, "Ooaspectz and aspect-oriented uml class diagrams for aspect-oriented software modelling (aosm)," *Ingeniería e Investigación*, vol. 33, no. 3, pp. 66–71, 2013.

[13] C. Vidal Silva, R. Villarroel, R. Schmal Simon, R. Saens, T. Tigero, and C. Del Rio, "Aspect-oriented formal modeling:(aspectz+ object-z)= ooaspectz," *Computing and Informatics*, vol. 34, no. 5, pp. 996–1016, 2016.

[14] A. O. AL-Zaghameem, "An aspect oriented programming framework to support transparent runtime monitoring of applications," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019.

[15] C. V. Silva, L. López, R. Schmal, R. Villarroel, M. Bustamante, and V. R. Sanchez, "Jpi uml software modeling," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 12, 2015.

[16] R. Laddad, *AspectJ in Action: Practical Aspect-Oriented Programming*. Greenwich, CT, USA: Manning Publications Co., 2003.

[17] T. Pender, *UML Bible*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 2003.

[18] D. Torre, Y. Labiche, M. Genero, and M. Elaasar, "A systematic identification of consistency rules for uml diagrams," *Journal of Systems and Software*, vol. 144, pp. 121–142, 2018.

[19] S. Kojarski and D. H. Lorenz, "Modeling aspect mechanisms: A top-down approach," in *Proceedings of the 28th International Conference on Software Engineering*, ser. ICSE '06. New York, NY, USA: ACM, 2006, pp. 212–221.

[20] F. F. Silveira, A. M. da Cunha, and M. L. Lisbôa, "A state-based testing method for detecting aspect composition faults," in *Computational Science and Its Applications – ICCSA 2014*, B. Murgante, S. Misra, A. M. A. C. Rocha, C. Torre, J. G. Rocha, M. I. Falcão, D. Taniar, B. O. Apduhan, and O. Gervasi, Eds. Cham: Springer International Publishing, 2014, pp. 418–433.

Artificial Potential Field Algorithm Implementation for Quadrotor Path Planning

Iswanto[1], Alfian Ma'arif*[2], Oyas Wahyunggoro[3], Adha Imam Cahyadi[4]

Department of Electrical Engineering[1]
Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia
Department of Electrical Engineering*[2]
Universitas Ahmad Dahlan, Yogyakarta, Indonesia
Corresponding Author*[2]
Department of Electrical Engineering and Information Technology[3,4]
Universitas Gadjah Mada, Yogyakarta, Indonesia

***Abstract*—Potential field algorithm introduced by Khatib is well-known in path planning for robots. The algorithm is very simple yet provides real-time path planning and effective to avoid robot's collision with obstacles. The purpose of the paper is to implement and modify this algorithm for quadrotor path planning. The conventional potential method is firstly applied to introduce challenging problems, such as not reachable goals due to local minima solutions or nearby obstacles (GNRON). This will be solved later by proposed modified algorithms. The first proposed modification is by adding virtual force to the repulsive potential force to prevent local minima solutions. Meanwhile, the second one is to prevent GNRON issue by adding virtual force and considering quadrotor's distance to goal point on the repulsive potential force. The simulation result shows that the second modification is best applied to environment with GNRON issue whereas the first one is suitable only for environment with local minima traps. The first modification is able to reach goals in six random tests with local minima environment. Meanwhile, the second one is able to reach goals in six random tests with local minima environment, six random tests with GNRON environment, and six random tests with both local minima and GNRON environment.**

Keywords*—Quadrotor; path planning; GNRON (Goal Non-reachable with Obstacles Nearby); artificial potential field; local minima

I. Introduction

The Artificial Potential field is a path planning algorithm for moving the robot from the initial to the goal point by the artificial potential field method found by Khatib [1]. It provides simple and effective motion planners for practical purpose [2].

The Artificial Potential field algorithm has been examined by several researchers for various applications such as mobile robots [3], [4], [5], wheelchair [6], underwater vehicles [7], [8], humanoid robots [9], walking robots [10], planetary rovers [11], autonomous sailboat [12], and bio-fish [13].

On the other side, quadrotors has become interesting topics on researches due to its functional uses such as surveillance robots [14] [15], autonomous target tracking [16], bridge crack detection [17], object inspection [18] and autonomous transportation systems [19]. However, there are only a few researches about potential field algorithm implementation on quadrotor path planning [27].

In 2016, Woods et al. conducted a study of potential field controls to be applied to the quadrotor with Ardrone type [20]. It proposed an extended APF-based algorithm to make quadrotor able to avoid collision with obstacles in aerial space. Unfortunately, the research had not provide any experiment with local minima or obstacle near goal environment. Meanwhile, at the same year, Mac et al using artificial potential field algorithm applied to quadrotor with Ardrone type to simulate it with gazebo simulator in the robot operation system (ROS) software [21]. The study conducts varies tests from known area to unknown obstacles. Yet, the XY-coordinates in the test are on positive axis only; there is no test for coordinates on negative axis. Therefore, there is no guarantee that the result will be the same if the initial, obstacle, and goal position changes to negative axis. Besides, every test was conducted only at the same starting positions.

The random force algorithm was studied by Lee et al to create new potential functions for to solve the problem of symmetrically aligned robot-obstacle-goal (SAROG) and the problem of local minima of the potential field algorithm [22]. By using the new potential function mobile robot can pass local minima and SAROG to reach the goal point. The local virtual target method was investigated by Zou & Zhu to create local virtual target attraction in the artificial potential field algorithm to avoid local minima so that the robot can reach the final point [23]. The concept of virtual obstacles was studied by Liu Chengqing et al. to modify the artificial potential field algorithm to close the local minima [24].

The research conducted and presented in this paper is different from the research previously mentioned. This paper presents the potential field algorithm in which the attractive and repulsive forces have been modified by using virtual potential so that when applied in the quadrotor path planning, the quadrotor will be able to avoid GNRON, local minima, and static and dynamical obstacles quickly.

II. Assumptions

These following assumptions are made to simplify the analysis of the research: Assumption 1, the shape, position, and the speed of the robot is known; Assumption 2, the position

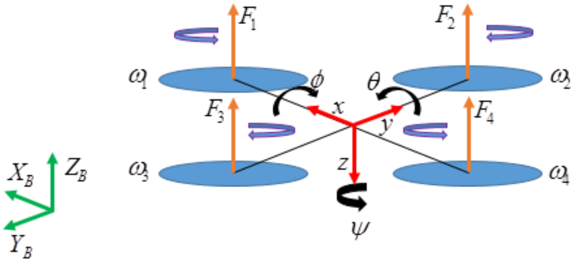


Fig. 1. Quadrotor Model

of the goal is known; Assumption 3, the static obstacles are polygonal-shaped with their positions are known.

III. Quadrotor Control Modeling

The quadrotor is an unmanned aircraft capable of moving upward by using the thrust generated by the rotation of the four propellers driven by electric motor rotation as shown in Fig. 1 which is the model of Mahony et al. [25]. As an unmanned aircraft, the quadcopter is often used as a means for monitoring, rescuing, and some military purposes. Like helicopters, the quadcopter has some advantages that are it can fly and land vertically in a narrow area. Also, a quadrotor can do rotation or movement easier.

Fig. 1 shows that the quadrotor has a body frame (x_b, y_b, z_b) in the global frame (x, y, z). The quadrotor movement is based on 6 degrees of freedom involving translational and rotation movements of the quadrotor position against the earth frame $[y\ z\ \phi\ \theta\ \psi]^T$. Based on the Fig. 1, the non-linear quadrotor model can be made, where $\delta = \left[\phi\ \theta\ \psi\ x\ y\ z\ \dot{\phi}\ \dot{\theta}\ \dot{\psi}\ \dot{x}\ \dot{y}\ \dot{z}\ \right]^T$. Thus the models of the quadrotor are as follows,

$$\dot{\delta}_1 = \dot{\phi} = \ddot{\phi} + \ddot{\theta} sin\phi tan\theta + \ddot{\psi} cos\phi tan\theta \tag{1}$$

$$\dot{\delta}_2 = \dot{\theta} = \ddot{\theta} cos\phi - \ddot{\psi} sin\phi \tag{2}$$

$$\dot{\delta}_3 = \dot{\psi} = \frac{sin\phi}{cos\theta}\ddot{\theta} + \frac{cos\phi}{cos\theta}\ddot{\psi} \tag{3}$$

where ϕ is the angle of the roll, θ is the angle of the pitch, ψ is the angle yaw, $\dot{\phi}$ is the angular velocity on the x-axis, $\dot{\theta}$ is the angular velocity on the y-axis, $\dot{\psi}$ is the angular velocity on the z-axis, $\ddot{\phi}$ is the angular acceleration on the x-axis, $\ddot{\theta}$ is the angular acceleration on the y-axis, $\ddot{\psi}$ is angular acceleration on the z-axis.

$$\dot{\delta}_4 = \dot{x} = \delta_{10} \tag{4}$$

$$\dot{\delta}_5 = \dot{y} = \delta_{11} \tag{5}$$

$$\dot{\delta}_6 = \dot{z} = \delta_{12} \tag{6}$$

where, $\dot{x}$ is the speed on the x-axis, $\dot{y}$ is the speed on the y-axis, and $\dot{z}$ is the speed on the z-axis. Then, the next equation of quadrotor model is as follows:

$$\dot{\delta}_7 = \ddot{\phi} = \frac{\tau_x}{I_x} - \frac{I_z - I_y}{I_x}\ddot{\theta}\ddot{\psi} \tag{7}$$

$$\dot{\delta}_8 = \ddot{\theta} = \frac{\tau_y}{I_y} - \frac{I_x - I_z}{I_y}\ddot{\phi}\ddot{\psi} \tag{8}$$

$$\dot{\delta}_9 = \ddot{\psi} = \frac{\tau_z}{I_z} - \frac{I_y - I_x}{I_z}\ddot{\phi}\ddot{\theta} \tag{9}$$

where, τ_x is the torque of the roll, τ_y is the torque of the pitch, τ_z is the torque of the yaw, I_x is the moment of inertia of the x-axis, I_y is the moment of inertia of the y-axis, and I_z is the moment of inertia of the z-axis. The non-linear model of quadrotor based on the is as follows:

$$\dot{\delta}_{10} = \ddot{x} = -\frac{1}{m}T(cos\phi sin\theta cos\psi + sin\phi sin\psi) \tag{10}$$

$$\dot{\delta}_{11} = \ddot{y} = -\frac{1}{m}T(cos\phi sin\theta sin\psi - sin\phi cos\psi) \tag{11}$$

$$\dot{\delta}_{12} = \ddot{z} = g - \frac{1}{m}T(cos\phi cos\theta) \tag{12}$$

where $\ddot{x}$ is the acceleration on the x-axis, $\ddot{y}$ is the acceleration on the y-axis o, $\ddot{z}$ is the acceleration on the z-axis, m is the quadrotor period, and T is the thrust. Some earlier researchers have designed a quadrotor control system by separating the non-linear quadrotor model into four subsystems as has been conducted by Ajmera & Sankaranarayanan [29]. There is four equations of the quadrotor control subsystem namely full-state feedback control system for roll τ_x, pitch τ_y, yaw τ_z, and thurst T subsystems.

Full-state feedback for the roll subsystem is

$$\tau_x = (k_1(x_{ref} - \delta_4) - k_2\dot{\delta}_4) + (k_3\delta_2 - k_4\dot{\delta}_2) \tag{13}$$

where k_1, k_2, k_3, and k_4 are the constants of the full-state feedback control gain for the roll subsystem.

Full-state feedback for the pitch subsystem is

$$\tau_y = (k_5(x_{ref} - \delta_5) - k_6\dot{\delta}_5) + (k_7\delta_1 - k_8\dot{\delta}_1) \tag{14}$$

where k_5, k_6, k_7, and k_8 are the constants of the full-state feedback control gain for the pitch subsystem.

Full-state feedback for the yaw subsystem is

$$\tau_z = -k_9(\delta_3 - \psi_{ref}) - k_{10}\dot{\delta}_3 \tag{15}$$

where k_9 and k_{10} are the constants of the full-state feedback control gain for the yaw subsystem.

Full-state feedback for thrust subsystems as follows:

$$T = k_{11}(z_{ref} - \delta_6) - k_{12}\dot{\delta}_6 \tag{16}$$

where, k_{11} and k_{12} are the constants of the full-state feedback control gain for the thrust subsystem.

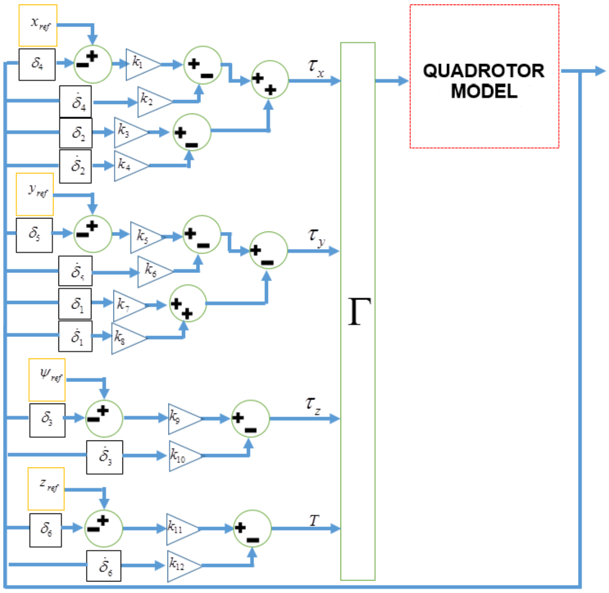


Fig. 2. PD Quadrotor Control

From the equation (13) - (16) about the full-state feedback control system, the PD algorithm control as shown in Fig. 2 can be designed. It can be seen that the there are four control subsystems namely roll, pitch, yaw, and thrust controls.

Fig. 2 shows that the full-state feedback control system has four outputs namely roll torque for the roll subsystem, pitch torque for the pitch subsystem, yaw torque for the yaw subsystem, and thrust torque for the thrust subsystem. The full-state feedback control system for the roll has four inputs namely the distance, and the speed on the x-axis, the angular position and the angular velocity of the roll. Full-state feedback control system for pitch has four inputs namely the distance on the y-axis, the speed on the y-axis, the angular position and the angular velocity of the pitch. The full-state feedback control system for yaw has two inputs namely the angular position and angular velocity of the yaw. The full-state feedback control system for the thrust has two inputs namely the distance of the z-axis and the speed on the z-axis. It is seen in the Fig. 2 that the torque relationship $\zeta = [\tau_x \quad \tau_y \quad \tau_z \quad T]^T$ is converted by using the matrix Γ to gain the rotor speed value $\Omega = [w_1 \quad w_2 \quad w_3 \quad w_4]$ on the quadrotor system shown in the following equation

$$\Omega = \Gamma\zeta \tag{17}$$

IV. Artificial Potential Field

The artificial potential field (APF) algorithm is one of the algorithms used in robot path planning in which its force F_{APF} is the sum of the attractive potential field F_{att} and the repulsive potential field F_{rep} as shown the following equation:

$$F_{APF} = F_{att} + F_{rep} \tag{18}$$

To apply the APF force to the robot, it is inserted into the linear speed equation on the kinematic robot [9]. The desired speed equation in Attractive APF force v_G^{att} is as follows:

$$v_G^{att}(x, y) = -\nabla U_{att}(x, y) \tag{19}$$

where the potential attractive equation is partially derived to the x and y-axis as follows:

$$v_x^{att}(x, y) = \frac{\partial U_{att}(x, y)}{\partial x} \tag{20}$$

$$v_y^{att}(x, y) = \frac{\partial U_{att}(x, y)}{\partial y} \tag{21}$$

The equation of the Khatib's potential attractive $U_{att}(x, y)$ of [1] is as follows:

$$U_{att} = \frac{1}{2}k_a((\delta_4 - x_{ref})^2 + (\delta_5 - y_{ref})^2) \tag{22}$$

where k_a is the potential attractive constant, δ_4, δ_5 is the position of the robot. (x_{ref}, y_{ref}) is the position of the goal point. The desired speed equation for the Attractive APF force v_G^{att} on the x and y-axis is as follows:

$$v_{G_x}^{att} = -k_a(\delta_4 - x_{ref}) \tag{23}$$

$$v_{G_y}^{att} = -k_a(\delta_5 - y_{ref}) \tag{24}$$

The desired speed equation in the repulsive force v_O^{rep} is

$$v_O^{rep}(x, y) = \nabla U_{rep}(x, y) \tag{25}$$

where the potential repulsive equations are partially derived to the x and y-axis as follows:

$$v_x^{rep}(x, y) = \frac{\partial U_{rep}(x, y)}{\partial_x} \tag{26}$$

$$v_y^{rep}(x, y) = \frac{\partial U_{rep}(x, y)}{\partial_y} \tag{27}$$

The equation of the Sfeir's et al potential repulsive $U_{rep}(x, y)$ of [30] is

$$U_{rep} = \begin{cases} \frac{1}{2}k_r(\frac{1}{\rho_O} - \frac{1}{r_O})^2 & \text{if} \rho_O \leq r_O \\ 0 & \text{if} \rho_O > r_O \end{cases} \tag{28}$$

where k_r is the potential repulsive constant, r_O is the distance limit of potential repulsive influence, and ρ_O is the closest distance between the robot and the obstacle as shown in Fig. 3.

The closest distance between the robot and the obstacle ρ_O is

$$\rho_O = \sqrt{x_{or}^2 + y_{or}^2} \tag{29}$$

where x_{or} is the difference of the distance between the robot and the obstacle on the x-axis, and y_{or} is the difference of the distance between the robot and the obstacle on the y-axis which the equation is as follows:

$$x_{or} = \delta_4 - x_O \tag{30}$$

$$y_{or} = \delta_5 - y_O \tag{31}$$

The following is the desired speed equation for the APF Repulsive force v_O^{rep} on the x and y-axes:

$$v_x^{rep} = \begin{cases} -k_r\left(1 - \dfrac{\rho_O}{r_O}\right)\dfrac{x_{or}}{\rho_O^3} & \text{if} \rho_O \leq r_O \\ 0 & \text{if} \rho_O > r_O \end{cases} \tag{32}$$

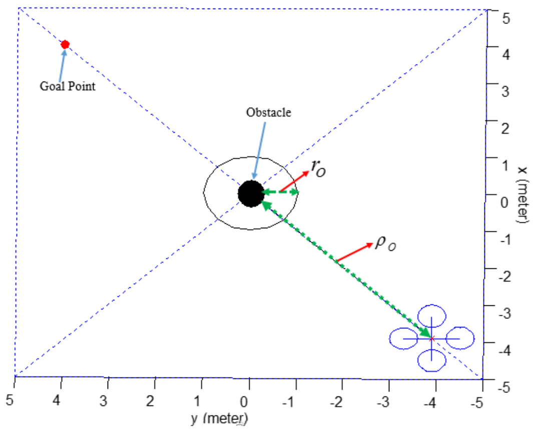


Fig. 3. The repulsive APF algorithm environment model

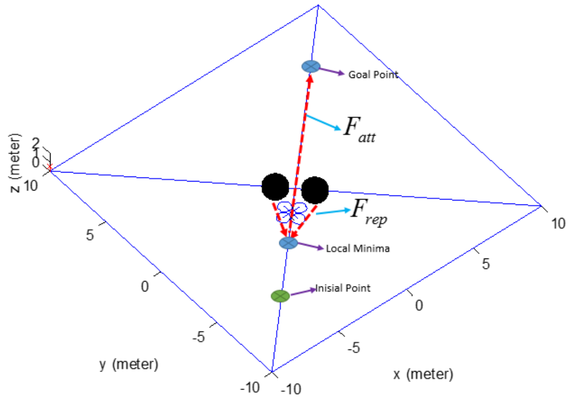


Fig. 5. Environment with local minima

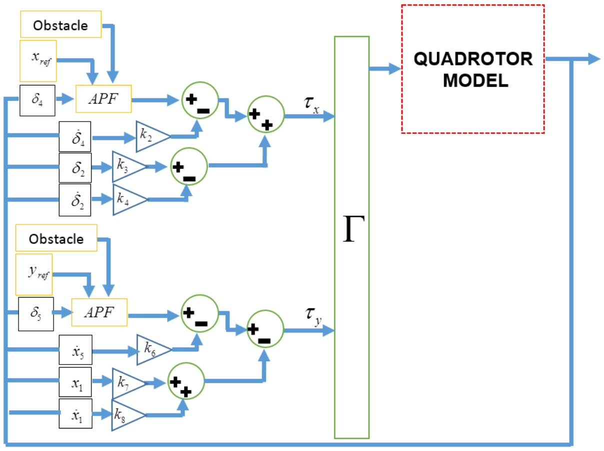


Fig. 4. Potential Field Algorithm for Quadrotor

$$v_y^{rep} = \begin{cases} -k_r \left(1 - \dfrac{\rho_O}{r_O}\right) \dfrac{y_{or}}{\rho_O^3} & \text{if} \rho_O \le r_O \\ 0 & \text{if} \rho_O > r_O \end{cases} \quad (33)$$

Thus, the speed equations of the x and y-axes in the APF are as follows:

$$v_x^{apf} = v_{G_x}^{att} + v_{O_x}^{rep} \quad (34)$$

$$v_y^{apf} = v_{G_y}^{att} + v_{O_y}^{rep} \quad (35)$$

The first problem found in this research is applying APF force to the quadrotor. This problem can be solved by substituting the speed equations on the x and y-axes of the APF into the full-state feedback control equations for the roll and pitch subsystems as shown in the following equations:

$$\tau_x = (v_x^{apf} - k_2\dot{\delta}_4 + (k_3\delta_2 - k_4\dot{\delta}_2)) \quad (36)$$

$$\tau_y = (v_y^{apf} - k_6\dot{\delta}_5 + (k_7\delta_1 - k_8\dot{\delta}_1)) \quad (37)$$

Based on the equation, the block diagram control of artificial potential field as shown in Fig. 4 can be designed. It is seen that there are two artificial potential fields namely the artificial potential field of the x and of the y-axes. The artificial potential field of the x-axis has three inputs: the obstacle data input of the x-axis, goal data input of the x-axis and robot data input of the x-axis, while the artificial potential field of the y-axis has three inputs: obstacle data input of the y-axis, goal data input of the y-axis and robot data input of the y-axis. The output of the artificial potential field algorithm is the vector speed of the x and y-axes. The speed vectors of the x and y-axes are inserted into the PD control system of the quadrotor.

V. GNRON AND LOCAL MINIMA PROBLEM

Artificial potential field algorithm is one of the algorithms in the robot path planning method used to reach the goal point and avoid obstacles by using the magnetic force method that is the attractive force to reach the goal point and repulsive force to avoid obstacles in an unknown environment. The problems appear when it is applied to the environment that has many obstacles such as local minima area as shown in Fig. 5. From the picture, it is seen that Quadrotor stops in local area minima which are formed by two obstacles placed in parallel which has a gap between them in an unknown environment.

In the APF algorithm, there is an area called local minima that is the area where the attractive potential field F_{att} and repulsive potential field F_{rep} has the same value resulting in a zero value of the artificial potential field F_{APF}. The environment that has the local minima area can be tested with the graph of the algorithmic force as shown in Fig. 6. It is seen that the two obstacles located at points (8.8,8) and (8,8.8) have a repulsive force to reject the quadrotor. It is also seen that in front of the two obstacles there is a basin indicating a local minimum due to the value of artificial potential field artificial force is zero. Point (0, 0) is the initial position of the robot and the value of the artificial potential field is large because the attractive force is proportional to the distance. Fig. 6 shows that point (12, 12) has zero value of artificial potential field force because it is the goal position of the robot.

In the APF algorithm there is an area called local minima which is the area where the attractive potential field F_{APF} and repulsive potential field F_{rep} have the same value resulting in a zero value of the artificial potential field F_{APF}. To avoid the area, the author proposed a modification of the artificial

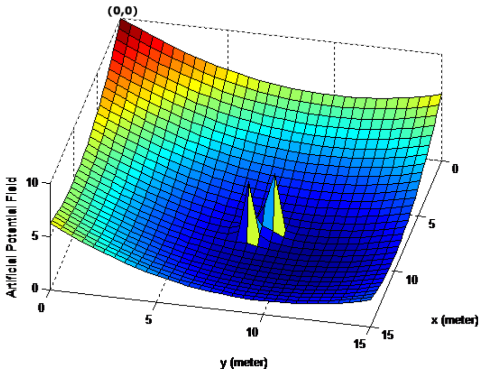


Fig. 6. Khatib's potential in local minima environment

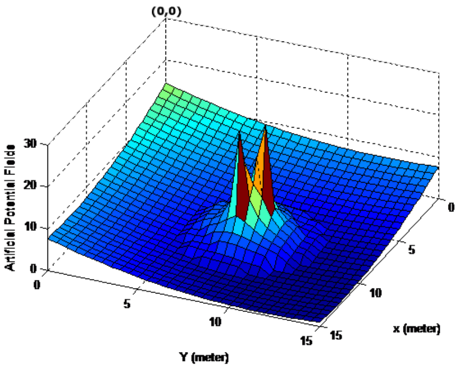


Fig. 7. The proposed APF potential in local minima environment

potential field algorithm by using virtual potential that is placed in front of the obstacle.

The equation of virtual potential $U_{vir}(x,y)$ is as follows:

$$U_{vir} = \begin{cases} \dfrac{k_v}{\rho_O} & \text{if} \rho_O \leq r_O \\ 0 & \text{if} \rho_O > r_O \end{cases} \tag{38}$$

where k_v is the potential repulsive constant.

The desired speed equation for the virtual potential force v_O^{vir} on the x and y-axes is as follows:

$$v_x^{vir} = \begin{cases} -k_v \dfrac{x_{or}}{\sqrt{(x_{or}^2 + y_{or}^2)^3}} & \text{if} \rho_O \leq r_O \\ 0 & \text{if} \rho_O > r_O \end{cases} \tag{39}$$

$$v_y^{vir} = \begin{cases} -k_v \dfrac{y_{or}}{\sqrt{(x_{or}^2 + y_{or}^2)^3}} & \text{if} \rho_O \leq r_O \\ 0 & \text{if} \rho_O > r_O \end{cases} \tag{40}$$

Thus, the desired speed equation for the repulsive artificial potential field force v_O^{rep} on the x and y axes is

$$v_x^{rep\_vir} = \begin{cases} v_x^{rep} - k_v \dfrac{x_{or}}{\sqrt{(x_{or}^2 + y_{or}^2)^3}} & \text{if} \rho_O \leq r_O \\ 0 & \text{if} \rho_O > r_O \end{cases} \tag{41}$$

$$v_y^{rep\_vir} = \begin{cases} v_y^{rep} - k_v \dfrac{y_{or}}{\sqrt{(x_{or}^2 + y_{or}^2)^3}} & \text{if} \rho_O \leq r_O \\ 0 & \text{if} \rho_O > r_O \end{cases} \tag{42}$$

An environment that has a local minima area can be eliminated by the equation of the repulsive force of the artificial potential field v_O on the x and y-axes. The environment using the algorithm is tested with the algorithmic force graph as shown in Fig. 7. It is seen that the two obstacles located at points (8.8, 8) and (8, 8.8) have a repulsive force to reject the quadrotor. It is also seen that in front of the two obstacles there is no basin indicating that there is no local minima. Point (0,0) is the initial position of the robot and the value of the artificial potential field is large because the attractive force is proportional with the distance. The Fig. 7 shows that point (12, 12) has zero value of artificial potential field force because it is the goal position of the robot.

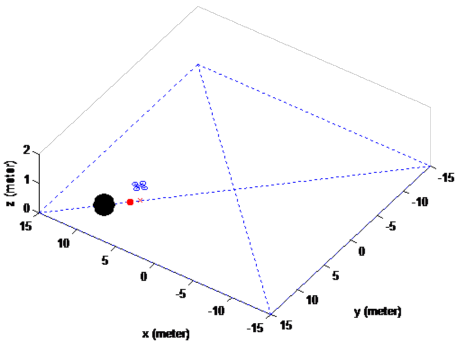


Fig. 8. The environment with GNRON

Besides local minima traps, there is another problem in the artificial potential field algorithm that is Goal Not Reachable when Obstacles are Nearby (GNRON). In this case the point is located near the obstacle causing the quadrotor is unable to reach the goal point as shown in Fig. 8. It is seen that the obstacle located at point (10, 10) and the goal point is located near the obstacle at point (8, 8). Fig. 8 shows that the quadrotor stops at the goal point and cannot reach the goal point due to the attractive potential value is the same as the repulsive potential.

GNRON is the area where the goal point is near the obstacle so that the quadrotor cannot reach the goal point because the attractive potential field F_{att} and repulsive potential field F_{rep} have the same value resulting in a zero value of the artificial potential force field F_{APF}. The goal near the obstacles is tested using the algorithmic force graph as shown in Fig. 9. It shows that the obstacles located at the point (12, 12) has a repulsive force to reject the quadrotor. It is seen that a basin near the obstacle is the robot goal point (10, 10) which has a zero value of artificial potential field. However, near the robot goal point, there is another basin at point (9, 9) which has zero value of artificial potential field causing the robot stop at that point and cannot reach the goal.

To reach the goal in a GNRON environment, the artificial potential field algorithm has been modified using the virtual potential force. The environment using the algorithm is tested by the algorithmic force graph shown in Fig. 10. Fig. 10 shows that the obstacle is placed at point (12, 12) has a repulsive force

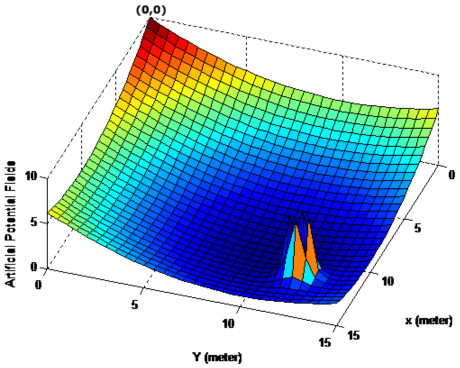


Fig. 9. Khatib's APF potential in the GNRON environment

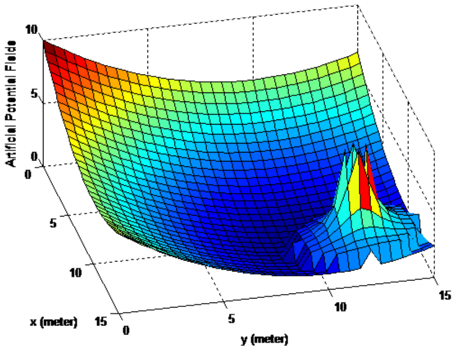


Fig. 10. The proposed APF potential in the GNRON environment

to reject the quadrotor. It is seen that there is a basin near the obstacle. The goal point of the robot at (10, 10) is in the basin and it has a zero value of artificial potential field. By using the modification of artificial potential field and virtual potential force, there is no basin, however, the robot cannot reach the goal because it is obstructed by the obstacle.

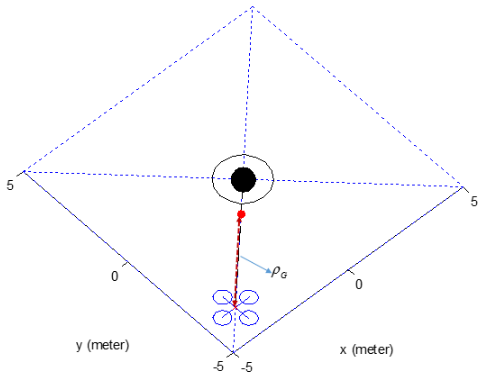


Fig. 11. GNRON environment

Based on the figure of the force graphic of the artificial potential field algorithm a shown in Fig. 10, the artificial potential field $U_{rep}(x,y)$ algorithm is modified as follows:

$$U_{rep} = \begin{cases} -\frac{1}{2}k_r\left(\frac{1}{\rho_O}-\frac{1}{r_O}\right)\rho_G^2 & \text{if}\rho_O \leq r_O \\ 0 & \text{if}\rho_O > r_O \end{cases} \tag{43}$$

where ρ_G is the distance of the robot to the goal point shown in Fig. 11. It is seen in the figure that the robotic goal point is located near the obstacle so that the quadrotor cannot reach the goal point, hence, the artificial potential field repulsive ρ_G is modified.

The distance between the robot and the goal ρ_G is

$$\rho_G = \sqrt{x_{Gr}^2 + y_{Gr}^2} \tag{44}$$

where x_{Gr} is the difference of the distance between the robot and the obstacle on the x-axis, and y_{Gr} is the difference of the distance between the robot and the obstacle on the y-axis which equation as follows:

$$x_{Gr} = \delta_4 - x_{ref} \tag{45}$$

$$y_{Gr} = \delta_5 - y_{ref} \tag{46}$$

The desired speed equation for the artificial potential field repulsive force v_O^{rep} on the x and y-axes is as follows:

$$v_x^{rep} = \begin{cases} -k_v\left(\frac{1}{\rho_O}-\frac{1}{r_O}\right)_2\frac{\rho_G^2}{\rho_O^3}x_{Or} & \text{if}\rho_O \leq r_O \\ -k_v\left(\frac{1}{\rho_O}-\frac{1}{r_O}\right) x_{Gr} & \text{if}\rho_O > r_O \end{cases} \tag{47}$$

$$v_y^{rep} = \begin{cases} -k_v\left(\frac{1}{\rho_O}-\frac{1}{r_O}\right)_2\frac{\rho_G^2}{\rho_O^3}y_{Or} & \text{if}\rho_O \leq r_O \\ -k_v\left(\frac{1}{\rho_O}-\frac{1}{r_O}\right) y_{Gr} & \text{if}\rho_O > r_O \end{cases} \tag{48}$$

From the new artificial potential field equation, the environment that has a goal point close to the obstacle is tested by algorithmic force graph shown in Fig. 12. In Fig. 12, it is seen that the obstacles located at the point (12, 12) have a repulsive force to reject the quadrotor. It shows that there is a basin in front of obstacles. The basin is the area where there is the robot goal point at (10, 10) which has an artificial potential field style equal to zero. By using artificial potential field algorithm that has been modified by the author, there is no more basin in front of the robot goal point that causes the robot stop so that the robot can reach the goal.

VI. Results and Discussion

The experiments presented in the paper are made with Peter Corke's quadrotor model [26], simulated in MATLAB simulation software. An algorithm then is applied to the quadrotor model to see its performances result.

There are three kinds of experiments represented in the paper. First, the original APF proposed by Khatib [1] is tested in MATLAB simulation. The algorithm is tested to local minima and GNRON environment. The performance results are analyzed to verify any found issues. Hence, the algorithm will be modified to solve the issues. There is two types of modified algorithms proposed in this paper. The first proposed

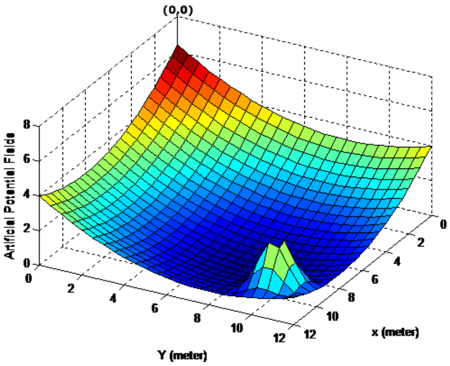


Fig. 12. The proposed new APF potential in the GNRON environment

modification is named as APF with virtual force. Meanwhile the second one is named as New APF. The proposed algorithms then are tested to environments which are similar to the previous experiment.

In the first experiment, the original APF was tested and applied to the quadrotor in an unknown environment where two static obstacles are located at points (-6.6) and (5, 5) as shown in Fig. 13. Fig. 13 shows that the quadrotor was tested five times by changing the starting and goal positions. The list of starting and goal positions are presented in Table I. In the first two tests, the obstacles, starting point, and destination point are located in a straight line.

The first test has local minima trap near the obstacles. Meanwhile, the second one does not have any local minima trap. The result shows that the quadrotor still able to avoid the local minima trap in the first test and reach goal points in all tests. However, the quadrotor does not move smoothly enough when avoiding the obstacle with local minima trap. In reality, the green line trajectory in the first test can be interpreted as the quadrotor moves back and forth in rapid movements. This can be bad for the quadrotor's safety.

TABLE I. START AND GOAL POSITIONS LIST IN FIRST EXPERIMENT ON FIG. 13

	Line Color	Start Point	Goal Point
Test1	Green	(-8,-8)	(10,10)
Test2	Blue	(0,-5)	(-10,10)
Test3	Red	(5,-8)	(-10,10)
Test4	Purple	(8,-8)	(-10,10)
Test5	Cyan	(8,0)	(-10,10)
Test6	Yellow	(0,8)	(-10,10)

Next, the original APF was applied to the quadrotor in an environment where the goal point is (-4,4) and a local minima area generated by two static obstacles located at point (-0.8, 0.8) and (0.8, 0.8) as shown in Fig. 14. The quadrotor is tested five times by changing the quadrotor starting point, and the list of starting points are shown in Table II. The result shows that the quadrotor reaches the goal point and able to avoid the local minima trap in all tests except in the second test.

The original APF then is applied to the quadrotor in a GNRON environment where there is a goal point located at point (2, 2) and a static obstacle located at point (0,0) as shown in Fig. 15. The quadrotor is tested six times by changing the quadrotor starting position as listed in Table III. The result shows that there is no quadrotor to reach the goal point in all tests.

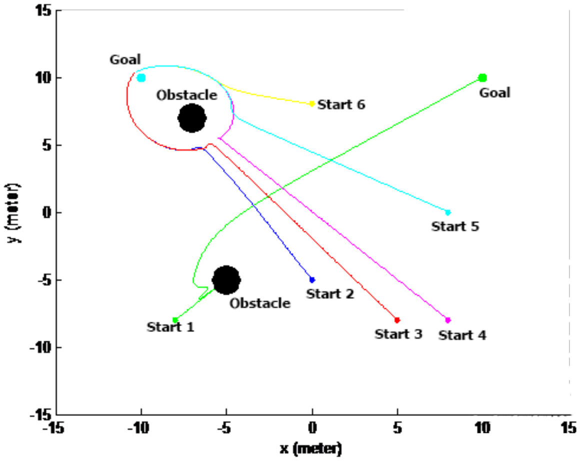


Fig. 13. Khatib's APF Algorithm Experiment

TABLE II. START POSITIONS LIST IN SECOND EXPERIMENT ON FIG. 14

	Line Color	Start Point	Reach Goal?
Test1	Red	(8,8)	No
Test2	Yellow	(0,8)	Yes
Test3	Blue	(-8,8)	Yes
Test4	Green	(-8,-8)	Yes
Test5	Purple	(8,-8)	Yes
Test6	Cyan	(8,0)	Yes

The quadrotor was not able to avoid the obstacle on the second test in the environment with local minima trap because of the straight line positions of start, obstacles, and goal points. Hence, the repulsive force is equal to attractive force or it can be assumed that the total forces happen to the quadrotor is equal to zero. There will be any force to move quadrotor toward the goal point and the quadrotor is trapped in local minima.

The quadrotor was not able to reach goal points located near obstacle in the environment with GNRON because its repulsive force is so much bigger than its attractive force in every test. Hence, the quadrotor is pushed to somewhere near goal points outside the repulsive force area of an obstacle.

TABLE III. START POSITIONS LIST IN THIRD EXPERIMENT ON FIG. 15

	Line Color	Start Point	Reach Goal?
Test1	Green	(-8,-8)	No
Test2	Purple	(8,0)	No
Test3	Blue	(-8,8)	No
Test4	Yellow	(0,8)	No
Test5	Red	(8,8)	No
Test6	Cyan	(8,0)	No

To overcome local minima trap issue, the total forces

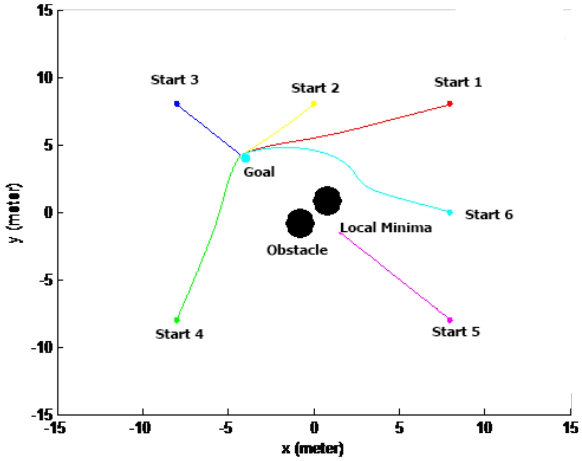


Fig. 14. Khatib's APF Algorithm Experimental in local minima environment

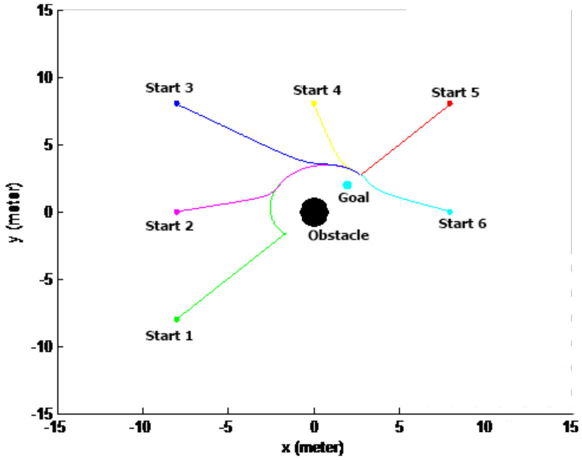


Fig. 15. Khatib's APF Algorithm Experiment in GNRON environment

happen to the quadrotor must not be equal to zero when it has not reached the goal point. To make it happened, a virtual force is added to the repulsive force. This equation contains a certain constant multiplied with the distance between the quadrotor and the obstacle. Hence, local minima can be eliminated.

Next experiment is to test the first proposed modified algorithm. The proposed APF with virtual force was applied to the quadrotor in an environment with local minima area generated from two static obstacles located at the point (-0.8, 0.8) and (0.8, -0.8) with a goal point (4, 4) as shown in Fig. 16. The quadrotor is tested six times by changing the initial position of the quadrotor as listed in Table IV. Fig. 16 shows that the quadrotor can avoid obstacles which have local minima area and moves toward the goal point in all tests. It also shows that the trajectory of the quadrotor is safe enough, and there is no back and forth rapidly movement.

Then, the APF with virtual force was applied to the quadrotor in a GNRON environment where there was a destination point at a point (2, 2) and a static obstacle at (0, 0) as shown in Fig. 17. It can be seen that the quadrotor tested six times the test by changing the quadrotor start position, as listed in Table V. From six times of experiments, it is seen that the quadrotor cannot reach the goal point in any test. Although it seems that the quadrotor reaches the goal point, it does not precisely reach the goal point. The quadrotor reaches the nearest point to the goal outside the repulsive area. However, when this result is compared to the original APF, the final destination point reached by the quadrotor is closer to the goal point.

TABLE IV. START POSITIONS LIST IN FOURTH EXPERIMENT ON FIG. 16

	Line Color	Start Point
Test1	Red	(8,8)
Test2	Yellow	(0,8)
Test3	Blue	(-8,8)
Test	Green	(-8,-8)
Test	Purple	(8,0)
Test	Cyan	(8,0)

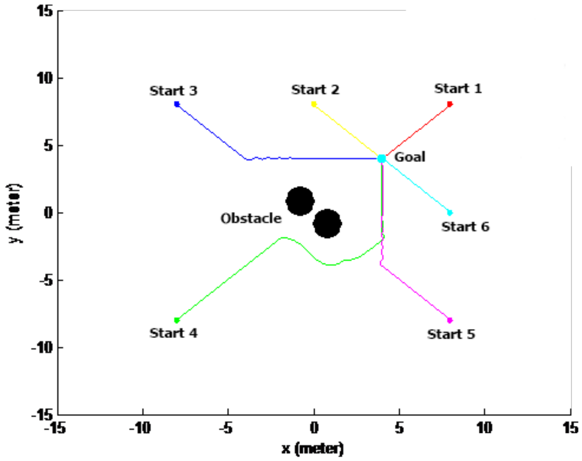


Fig. 16. Virtual APF Algorithm Experiment in local minima environment

TABLE V. START POSITIONS LIST IN FIFTH EXPERIMENT ON FIG. 17

	Line Color	Start Point
Test1	Green	(-8,-8)
Test2	Purple	(8,0)
Test3	Blue	(-8,8)
Test4	Yellow	(0,8)
Test5	Red	(8,8)
Test6	Cyan	(8,0)

Regarding previous experiment result's analysis, GNRON environment issue happens because there is no consideration of its distance to goal point in repulsive force. The repulsive force should be smaller when the quadrotor's distance toward the goal is smaller too. Therefore, the repulsive force should be modified by considering the quadrotor's distance to the goal point. This second proposed modification algorithm is called New APF.

The next experiment is the application of the New APF to the quadrotor in an environment where goal point is located

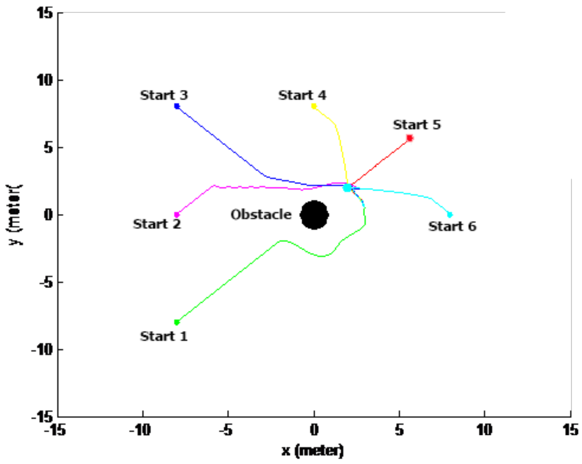


Fig. 17. The Virtual Force APF Algorithm Experiment in GNRON environment

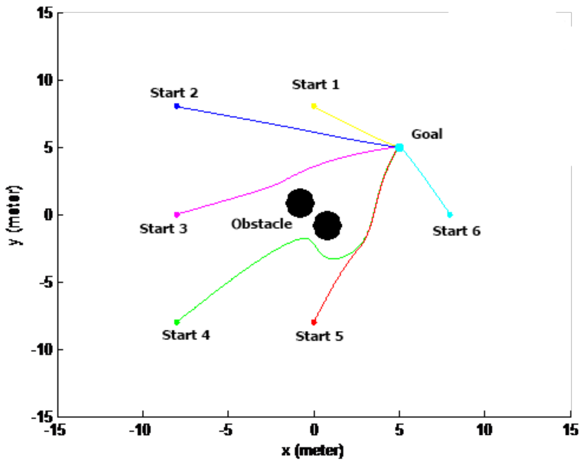


Fig. 18. New APF Algorithm Experiment in Local Minima Environment

at point (5, 5) and local minima generated from two static obstacles at point (-0.8, 0.8) and (0.8, -0.8) as shown in Fig. 18. The quadrotor is tested six times by changing the initial position of the quadrotor as listed in Table VI. It can be seen in Fig. 18 that the quadrotor avoids the area of local minima generated from two adjacent obstacles and moves toward the goal point in the first test. This result also provides better trajectories than two previous experiments. The trajectory is the smoothest among algorithms. The quadrotor can reach the goal point in all tests.

TABLE VI. Start Positions List in Sixth Experiment on Fig. 18

	Line Color	Start Point
Test1	Yellow	(0,8)
Test2	Blue	(-8,8)
Test3	Purple	(8,0)
Test4	Green	(-8,-8)
Test5	Red	(8,8)
Test6	Cyan	(8,0)

Next, the New APF was applied to the quadrotor in GNRON environment where the goal point is located at point (2, -2) and a static obstacle is located at point (0,0) as shown in Fig. 19. The Fig. 19 shows that the quadrotor is tested six times by changing the initial position of the quadrotor, as seen in Table VII.

From Fig. 19, it is seen that the quadrotor can avoid obstacles and can go to the goal point located near the obstacle. The quadrotor reaches targets in all tests.

In further analysis, some improvements can be seen by comparing this result with previous experiment's (APF with virtual force) result. Although it seems that it is safer to be passed by the quadrotor in previous experiment result, the trajectory was not optimal due to its unoccupied space between obstacle and the trajectory. In other words, the trajectory in previous result was safe but was not optimal in its mileage. However, this experiment trajectories are optimal by the mileage yet not really safe to be passed by the quadrotor.

TABLE VII. Start Positions List in Seventh Experiment on Fig. 19

	Line Color	Start Point
Test1	Yellow	(0,8)
Test2	Blue	(-8,8)
Test3	Purple	(8,0)
Test4	Green	(-8,-8)
Test5	Red	(8,8)
Test6	Cyan	(8,0)

Last but not least, the New APF was applied to the quadrotor in an environment with local minima trap and GNRON. The goal point is located at point (2, -2) and a local minima generated from two static obstacles located at the point (-0.8, 0.8) and (0.8, -0.8). The quadrotor is tested six times by changing the initial position of the quadrotor as listening in Table VIII. The result is shown as in Fig. 20. It is proved that the quadrotor can avoid local minima traps and to reach the goal point near the obstacle. The quadrotor reaches the target in all tests.

Uniquely, the trajectories made by this experiment result is not only safe enough to be passed by the quadrotor but also provides optimal trajectories by considering the mileage. The trajectories are safe because there is still some distance with the obstacle and the trajectories are smooth enough (no sharp turns or sudden maneuvers). However, the distance with the obstacle is not too far so it is still optimal enough.

TABLE VIII. Start Positions List in Eighth Experiment on Fig. 20

	Line Color	Start Point
Test1	Yellow	(0,8)
Test2	Blue	(-8,8)
Test3	Purple	(8,0)
Test4	Green	(-8,-8)
Test5	Red	(8,8)
Test6	Cyan	(8,0)

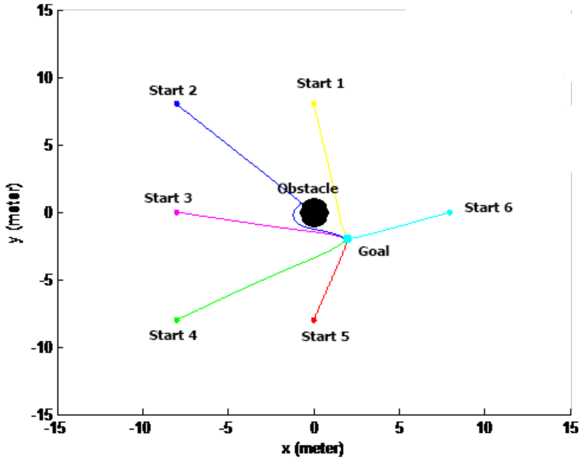


Fig. 19. New APF Algorithm Experiment in GNRON Environment

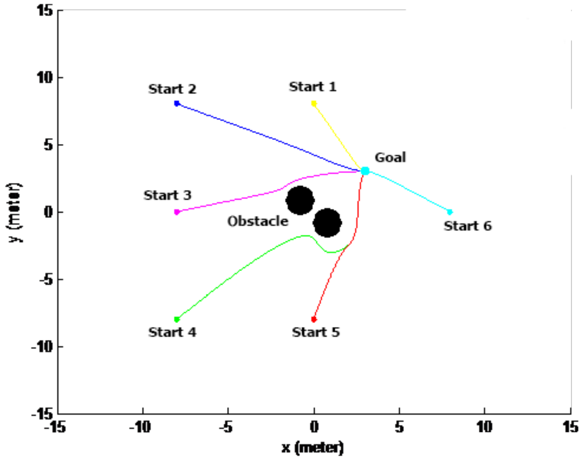


Fig. 20. New APF Algorithm experiment in local minima and GNRON

VII. Conclusion

This paper presents modified Khatib's potential field algorithm applied to the quadrotor. There are problems related to the Khatib's algorithm that is the attractive force to the goal located near the obstacle is not functioned because there is a repulsive force from the obstacle. The other problem is that the repulsive force used to avoid the obstacle make the quadrotor stop. To overcome the problem, the attractive and repulsive forces of the Khatib's potential field algorithm was modified. With the modification of the algorithm, the quadrotor can quickly avoid the static and dynamic obstacles and the local minima compared to the Khatib's algorithm.

Appendix

New Artificial Potential Field

The new artificial potential field $U_{rep}(x, y)$ algorithm is modified as follows. The APF because of the position is,

$$U_{rep} = U_{rep\_1} U_{rep\_2}$$

where,

$$U_{rep\_1} = -\frac{1}{2} k_r \left(\frac{1}{\rho_O} - \frac{1}{r_O} \right)^2$$
$$U_{rep\_2} = \rho_G^2$$

then,

$$U_{rep} = -\frac{1}{2} k_r \left(\frac{1}{\rho_O} - \frac{1}{r_O} \right)^2 \rho_G^2$$

While, the APF because of the velocity is,

$$\begin{aligned} v_x^{rep}(x,y) &= \frac{\partial U_{rep\_1}(x,y)}{\partial \rho_O} \frac{\partial \rho_O}{\partial x} U_{rep\_2}(x,y) \\ &+ \partial U_{rep\_1}(x,y) \frac{U_{rep\_2(x,y)}}{\partial \rho_G} \frac{\partial \rho_G}{\partial x} \\ &= -k_v \left(\frac{1}{\rho_O} - \frac{1}{r_O} \right) \frac{1}{\rho_O^2} \frac{x_{Or}}{\rho_O} \rho_G^2 \\ &+ k_v \left(\frac{1}{\rho_O} - \frac{1}{r_O} \right)^2 \rho_G \left(\frac{x_{Gr}}{\rho_G} \right) \end{aligned}$$

$$\begin{aligned} v_y^{rep}(x,y) &= \frac{\partial U_{rep\_1}(x,y)}{\partial \rho_O} \frac{\partial \rho_O}{\partial y} U_{rep\_2}(x,y) \\ &+ \partial U_{rep\_1}(x,y) \frac{U_{rep\_2(x,y)}}{\partial \rho_G} \frac{\partial \rho_G}{\partial y} \\ &= -k_v \left(\frac{1}{\rho_O} - \frac{1}{r_O} \right) \frac{1}{\rho_O^2} \frac{y_{Or}}{\rho_O} \rho_G^2 \\ &+ k_v \left(\frac{1}{\rho_O} - \frac{1}{r_O} \right)^2 \rho_G \left(\frac{y_{Gr}}{\rho_G} \right) \end{aligned}$$

Acknowledgment

The authors would like to thank to the editor and blind reviewer that give the advice and suggestion to significance improvement of the paper.

References

[1] O. Khatib, “Real-time obstacle avoidance for manipulators and mobile robots,” in Proceedings. 1985 IEEE International Conference on Robotics and Automation, vol. 2, pp. 500-505.

[2] Montiel, O., Orozco-Rosas, U., & Sepúlveda, R., ”Path planning for mobile robots using Bacterial Potential Field for avoiding static and dynamic obstacles,” Expert Systems with Applications, vol. 42 no. 12, pp. 5177-5191, 2015.

[3] Triharminto, H. H., Wahyunggoro, O., Adji, T. B., & Cahyadi, A. I., ”An Integrated Artificial Potential Field Path Planning with Kinematic Control for Nonholonomic Mobile Robot,” International Journal on Advanced Science, Engineering and Information Technology, vol. 6, no. 4, pp. 410-418, 2016.

[4] Triharminto, H. H., Wahyunggoro, O., Adji, T. B., Cahyadi, A., & Ardiyanto, I., ”Local Information using Stereo Camera in Artificial Potential Field based Path Planning,” IAENG International Journal of Computer Science, vol. 44, no. 3, 2017.

[5] O. Montiel, U. Orozco-Rosas, and R. Sepúlveda, “Path planning for mobile robots using Bacterial Potential Field for avoiding static and dynamic obstacles,” Expert Syst. Appl., vol. 42, no. 12, pp. 5177-5191, Jul. 2015.

[6] Onyango, S. O., Hamam, Y., Djouani, K., Daachi, B., & Steyn, N., ”A driving behaviour model of electrical wheelchair users,” Computational intelligence and neuroscience, 2016.

[7] G. Yun, W. Zhiqiang, G. Feixiang, Y. Bo, and J. Xiaopeng, “Dynamic Path Planning for Underwater Vehicles Based on Modified Artificial Potential Field Method,” in 2013 Fourth International Conference on Digital Manufacturing & Automation, 2013, pp. 518-521.

[8] S. F. J., R. A. F., V. A. Sebastián, and A. G. G., “Artificial potential fields for the obstacles avoidance system of an AUV using a mechanical scanning sonar,” in 2016 3rd IEEE/OES South American International Symposium on Oceanic Engineering (SAISOE), 2016, pp. 1-6.

[9] Q. Zhong, J. Zhao, and C. Tong, “Tracking for humanoid robot based on Kinect,” in 2014 International Conference on Mechatronics and Control (ICMC), 2014, pp. 1191–1194.

[10] H. Igarashi and M. Kakikura, “Path and posture planning for walking robots by artificial potential field method,” in IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA ’04. 2004, 2004, pp. 2165-2170 Vol.3.

[11] D. L. Sancho-Pradel and C. M. Saaj, “Assessment of Artificial Potential Field methods for navigation of planetary rovers,” in 2009 European Control Conference (ECC), 2009, pp. 3027-3032.

[12] F. Plumet, H. Saoud, and Minh-Duc Hua, “Line following for an autonomous sailboat using potential fields method,” in 2013 MTS/IEEE OCEANS - Bergen, 2013, pp. 1–6.

[13] Kong Feng, Xie Chaoping, and Wang Zijian, “Appication of a mixed method in bio-fish path planning,” in 2010 2nd International Conference on Education Technology and Computer, 2010, pp. V5-17-V5-19.

[14] K. W. Weng and M. S. b. Z. Abidin, “Design and Control of a Quad-Rotor Flying Robot For Aerial Surveillance,” in 2006 4th Student Conference on Research and Development, 2006, pp. 173–177.

[15] K. Mohta, M. Turpin, A. Kushleyev, D. Mellinger, N. Michael, and V. Kumar, “QuadCloud: A Rapid Response Force with Quadrotor Teams,” Springer, Cham, 2016, pp. 577–590.

[16] K. Boudjit and C. Larbes, “Detection and Implementation Autonomous Target Tracking with a Quadrotor AR.Drone,” in Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics, 2015, pp. 223–230.

[17] K. Hausman, J. Müller, A. Hariharan, N. Ayanian, and G. S. Sukhatme, “Cooperative Control for Target Tracking with Onboard Sensing,” Springer, Cham, 2016, pp. 879–892.

[18] K. Mathe, L. Busoniu, L. Barabas, C.-I. Iuga, L. Miclea, and J. Braband, “Vision-based control of a quadrotor for an object inspection scenario,” in 2016 International Conference on Unmanned Aircraft Systems (ICUAS), 2016, pp. 849-857.

[19] G. Loianno and V. Kumar, “Cooperative Transportation Using Small Quadrotors Using Monocular Vision and Inertial Sensing,” IEEE Robot. Autom. Lett., vol. 3, no. 2, pp. 680-687, Apr. 2018.

[20] A. C. Woods, H. M. Lay, and Q. P. Ha, “A novel extended potential field controller for use on aerial robots,” in 2016 IEEE International Conference on Automation Science and Engineering (CASE), 2016, pp. 286-291.

[21] T. T. Mac, C. Copot, A. Hernandez, and R. De Keyser, “Improved potential field method for unknown obstacle avoidance using UAV in indoor environment,” in 2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI), 2016, pp. 345-350.

[22] Lee, J., Nam, Y., Hong, S., & Cho, W, ”New potential functions with random force algorithms using potential field method,” Journal of Intelligent & Robotic Systems, 2012, 66(3), 303-319.

[23] Zou, X. Y., & Zhu, J. (2003). ”Virtual local target method for avoiding local minimum in potential field based robot navigation,” Journal of Zhejiang University-Science A, 4(3), 264-269.

[24] Chengqing, L., Ang, M. H., Krishnan, H., & Yong, L. S. (2000, April). ”Virtual obstacle concept for local-minimum-recovery in potential-field based navigation,” In Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065) (Vol. 2, pp. 983-988). IEEE.

[25] R. Mahony, V. Kumar, and P. Corke, “Multirotor Aerial Vehicles: Modeling, Estimation, and Control of Quadrotor,” IEEE Robot. Autom. Mag., vol. 19, no. 3, pp. 20-32, Sep. 2012.

[26] Corke, Peter. ”Robotics, vision and control: fundamental algorithms in MATLAB® second”, completely revised. Vol. 118. Springer, 2017.

[27] Chen, Y. B., Luo, G. C., Mei, Y. S., Yu, J. Q., & Su, X. L., ”UAV path planning using artificial potential field method updated by optimal control theory,” International Journal of Systems Science, 47(6), 1407-1420.

New Transport Layer Security using Metaheuristics and New Key Exchange Protocol

Mohamed Kaddouri[1], Mohammed Bouhdadi[2]
LMPHE Laboratory Mohammed V University,
Faculty of Sciences Rabat
Rabat, Morocco. BP1014 RP

Zakaria Kaddouri[3], Driss Guerchi[4], Bouchra Echandouri[5]
Department of Computer Science,
Mohammed V University Abu Dhabi
Abu Dhabi, United Arab Emirates. P.O. 106621

***Abstract*—The easiness of data transmission is one of the information security flaws that needs to be handled rigorously. It makes eavesdropping, tampering and message forgery by malicious more simple. One of the protocols developed to secure communication between the client and the server consists of using Transport Layer Security (TLS). TLS is a cryptographic protocol that allows encryption using record protocol, authentication and data integrity. In this paper, a new TLS version is proposed, named Transport Layer Security with Metaheuristics (TLSM), which is based on a recently designed metaheuristic symmetric ciphering technique for data encryption, combined with hash function SHA-SBOX and a new method for private key exchange. Compared to the existing TLS versions, the suggested protocol outperform all of them in terms of level of security of the encrypted data, key management and execution time.**

Keywords*—*Transport Layer Security (TLS); metaheuristic; symmetric ciphering algorithm; private key exchange; hash function

I. Introduction

Computer communication involving online transactions and payment in exchange for goods and services grew over the last decade drastically. However, an untrustworthy e-commerce website makes the costumers not having sufficient trust. One of the protocols that is widely employed to secure these transactions by providing authentication and encryption is the Transport Layer Security (TLS) protocol [12]. In this protocol, the encryption is used to prevent the interception of sensitive data, such as credit card numbers and account passwords.

TLS is a cryptographic protocol that provides end-to-end secure communication for different types of applications [12]. It was adopted by the IETF and specified as an RFC standard. It is the most widely used protocol between a client and a server. It is composed of two main parts: the Handshake Protocol and the Record Protocol [1]. The Handshake Protocol is in charge of key establishment employed to cipher data in the Record Protocol [13]. The simplicity of the handshake is due to its use of public key cryptography, which allows the negotiation of a shared secret key over a risky channel and without prior knowledge between the client and the server.

As a part of the TLS, the handshake protocol also allows the authentication of the presumed identity. The Record Protocol ensures a secure channel for the management of data delivery. The TLS protocol also provides its own data framing and authenticating method [12]. Despite all these security measures, the last version of the TLS protocol has suffered recently from several attacks [11, 14]. Namely, Denial of Service Attacks that attempt to saturate the server with SYN queries during the Handshake. Also, there are other attacks, such as SWEET32 (CVE-2016-2183, CVE-2016-6329)[15], DROWN (CVE-2016-0800)[16] and POODLE (CVE-2014-0160) [17], that are mainly related to weaknesses detected during the encryption.

According to the recent work presented in [2], several Bleichenbacher oracle attacks have shown some vulnerabilities in the TLS1.3 since they succeeded in the cryptanalysis of the RSA encrypted message [3].

The TLS uses RSA to exchange the secret key. This key is shared and decided by the client, then encrypted by the (server) public key before being sent to the server [1]. However, the process of key exchange using RSA [3] has several drawbacks, such as the slowness of any new connection [14].

In order to increase the speed of new connections, in this paper the use of the recently developed symmetrical metaheuristics ciphering technique [4] is suggested in the most sensitive phase of TLS protocol. This technique helps secure messages and private keys (session keys) exchanged between the client and the server. The simulation results show better execution time performance and higher security robustness.

The rest of this paper is organized as follows: in Section 2, related work are presented. Then in Section 3, the proposed solution TLSM are describe in details. Section 4 discusses the performance evaluation and security analysis of this approach. The last section concludes this paper and presents future works.

II. Related work

As mentioned in the introduction, TLS is a well-known and employed cryptographic protocol. The majority of security protocols do not use a mechanism about how to distribute secret keys. However, asymmetric cryptography, in particular RSA, has been declared much resource consuming for limited devices. Further, the last proposed versions of TLS suffer from some vulnerabilities that lead to attack success [2].

In Eronen et al. paper [5], the authors specified a set of cipher-suits for the Transport Layer Security protocol that support symmetric pre-shared keys based authentication. These processes are very slow because of the use of the Diffie-Hellman key exchange in authentication with the pre-shared key. The server and the client are authenticated with asymmetric encryption using pre-shared key.

The work Sizzle [6], a tinny version of TLS, was improved using Elliptic Curve Cryptography public key cryptosystem. It helps secure an end to end communication for devices with tight computational memory and energy constraints. However, it loses some security in its process.

Furthermore, in the paper [7], the authors show weaknesses of TLS. They presented a recovery attack against TLS when RC4 is employed in encryption. Their attack is based on the statistical analysis of RC4.

III. New Transport Layer Security Protocol using Metaheuristics

A. Background

1) Transport Layer Security : Over the past years, the Internet Engineering Task Force has been continuously working on developing one of its most important security protocols: the Transport Layer Security (TLS). It is a cryptographic protocol that ensures web security through encryption and authenticity of https website. It is supported by mainly two protocols: The Handshake Protocol and the Record Protocol [1].

The Handshake Protocol is conceived to negotiate a key security between a client and server. The handshake process is started by the client sending the message: $'ClientHello'$ including a random number and cipher suites. To respond to this message, the server sends a random number, a chosen cipher suit and a server certificate. The server's certificate is then authenticated by the client who generates a secret key. This secret key is encrypted before transmission by the client using the server public key. Once received by the server, the encrypted secret key is decrypted and used in the subsequent encryption steps [8].

The Record Protocol guarantees a secure channel for the management of data delivery. It starts by splitting the data into series of blocks. Afterwards, these blocks are compressed. Then, a message authentication code (MAC) is applied to these compressed blocks using the shared secret key. Thereafter, this block is encrypted with a symmetric encryption algorithm [8].

2) Symmetrical Metaheuristics Ciphering Approach : Symmetrical metaheuristics ciphering is a new cipher system using Vigenere and metaheuristics [4]. The aim of the use of Vigenere encryption is to maximize the confusion and the use of metaheuristics helps generate a robust secret Meta-key.

At the beginning of the encryption process, Vigenere algorithm is applied. Thereafter, a random key is generated. The ASCII number of every key's character will be added. The initial cipher (Vigenere cipher) is created and is given in table. Thereafter to every value is assigned a list of coordinates of the plain value. Shuffling begins by permutation, then the best solution is chosen using metaheuristic algorithms and evaluated by the evaluation function.

The final cipher is formed using the new list and each value is assigned to the coordinates stored in the beginning.

3) SHA-SBOX: SHA-SBOX [9] is an iterated hash function, inspired from SHA hash function, where the compression function involves permutation and substitution. Further, the Boolean functions Ch and Ma, used in SHA family, have been replaced by the substitution and permutation functions FPS. This FPS function takes three blocks of 32 bits, concatenates them and splits them into two blocks of 48 bits each. The two 48-bit blocks are then subject to permutations P1 and P2, respectively [9].

A substitution functions are then applied to the two 48-bit blocks producing hence two blocks of 32 bits each. The substitution functions are the well-known S-boxes provided in the DES encryption algorithm to ensure a good confusion. Thereafter, a modular addition is applied on the two 32-bit blocks.

B. TLSM Description

1) TLSM handshake protocol: The proposed TLSM protocol consists of several steps (Fig. 1). It uses both asymmetric and symmetric encryptions. The client and the server begin with a negotiation of the employed algorithms and the adopted an exchange of the key. In the handshake protocol, which is the most essential phase of establishing a secure connection, the server and the client exchange information are used to define the connection properties.

a. Step 1: Client Hello In the first step, the client sends a $\backslash ClientHello$” message to the server. This message includes the following list of information:
    - ○ Client Version of the TLSM with the sequence of supported algorithms.
    - ○ Client Time-Date, a 4-byte date representing the current date and time of the client (in epoch format).
    - ○ Session ID employed for the connection. The server searches for previous sessions, if this session is not empty, it remains in the current session.
    - ○ Compression methods employed for compressing TLSM packets.
    - ○ Cipher Suites (the combinations of cryptographic methods). It contains one cryptographic algorithm for namely: key exchange, data encryption and authentication. The proposed TLSM cipher suite is: $TLSM\_ECDHE\_ECDSA\_WITH$ $\_SMC\_SHASBOX$. It consists of the following information:
        - ▪ TLSM :Transport Layer Security with Metaheuristics.
        - ▪ ECDHE (Elliptic curve Diffie–Hellman): indicates the key exchange algorithm being employed.
        - ▪ ECDSA (Elliptic Curve Digital Signature Algorithm): indicates the authentication algorithm being employed.
        - ▪ SMC (Symmetrical Metaheuristic Ciphering): the ciphering algorithm.
        - ▪ SHA-SBOX: indicates the message authentication algorithm that is used to authenticate a message.
        - ▪ Compression methods: TLS compression methods, the data will be compressed before ciphering it.

b. Step 2: Server Hello The server replies with a $\backslash ServerHello$” when it receives $\backslash ClientHello$”

from the client side, that contains the proposed and selected options during the $\backslash ClientHello$” or a failure message.

- ○ Server Version is TLSM protocol.
- ○ Server Time-date indicates the current date and time of the client.
- ○ Session ID serves for a new session and resume sessions already open.
- ○ Cipher Suites. The server employs the Cipher Suites sent in the $\backslash ClientHello$”.
- ○ Compression Methods. The server will use the Compression Methods sent in the $\backslash ClientHello$”.

c. Step 3: Server Certificate The server, at this step, sends a signed TLSM certificate containing its public key to prove its identity to the client.

d. Step 4: Client Certificate (Optional) The client must provide his signed certificate, in case of the server asks the client to be authenticated with his certificate.

e. Step 5: Server Key Exchange This message will be sent to the server in case the certificate provided by the server is not sufficient for the client to exchange the symmetrical encryption Key.

f. Step 6: Server Hello Done This message is be sent to the client to affirm that the $\backslash ServerHello$” message is completed.

g. Step 7: Client Key Exchange Once the $\backslash HelloDone$” message is received from the server, the Client send its $\backslash KeyExchange$” message. Then, the $\backslash ClientKeyExchange$” is sent in case the server asks for a client certificate. Thereafter, the client generates the Meta-secret key. It is worth mentioning that before transmitting the Meta-secret key to the server, the client ciphers this key using the server’s public key. The asymmetric encryption is employed for the Meta-secret key exchange. Once the server receives the Meta-secret key, it employs its private key to decrypt it. Thereafter, for the rest of the communication, the latter is used by both the client and the server to encrypt and decrypt the data respectively.

h. Step 8: Client Change Cipher Spec Once the $\backslash ClientKeyExchange$” is finished all data communication between the client and the server is secure. The protocol $\backslash ChangeCipherSpec$” is employed to change the encryption.

i. Step 9: Client Handshake Finished This message is sent when the server receives the last message of the handshake process from the client.

j. Step 10: Server Change Cipher Spec All data sent communication, at this step, in the server side is secure.

k. Step 11: Server Handshake Finished This message will be sent when the client receives the last message of the handshake process from the server.

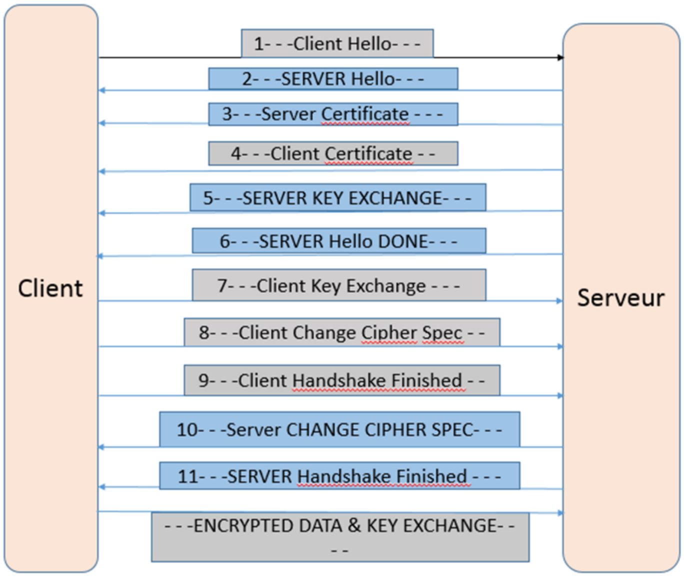

Client
Serveur
1- - -Client Hello- - -
2- - -SERVER Hello- - -
3- - -Server Certificate - - -
4- - -Client Certificate - -
5- - -SERVER KEY EXCHANGE- - -
6- - -SERVER Hello DONE- - -
7- - -Client Key Exchange - - -
8- - -Client Change Cipher Spec - -
9- - -Client Handshake Finished - -
10- - -Server CHANGE CIPHER SPEC- - -
11- - -SERVER Handshake Finished - - -
- - -ENCRYPTED DATA & KEY EXCHANGE- - - -

Fig. 1. TLSM Handshake Protocol.

2) *TLSM record protocol :* The TLSM consists of the following steps (Fig. 2):

a. Step 1 (Data): Application data blocks are split to several blocks with the same size.

b. Step 2 (Fragment): The first fragment is subject to compression, producing the output C(F1).

c. Step 3 (Hash Function): the digit termed E(F1) of C(F1) is calculated using the SHA-SBOX hash function [9].

d. Step 4 (Symmetric Encrypt & Meta-key Exchange): E(F1) is appended at the tail of C(F1), the result is encrypted by the symmetrical metaheuristics ciphering [4] using the first meta-key obtained from the handshake protocol , hence resulting in the encrypted data.

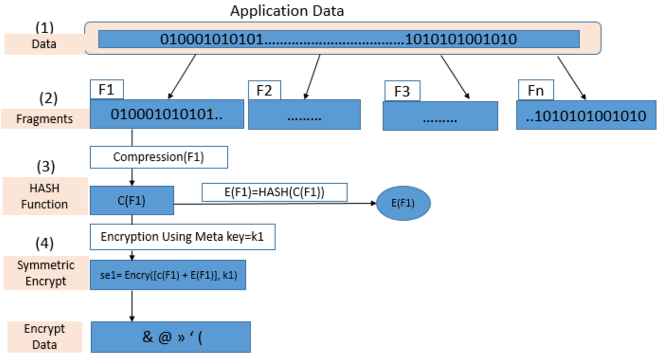


Fig. 2. TLSM record protocol.

3) *TLSM Key Exchange and Encryption Process :* Before presenting the operations of the key exchange and encryption processes, the following adopted parameters should be defined:

- N: length of the whole data

- N_f: length of each data frame F
- L_{km}: length (in bits) of the Metaheuristic key $k_{m,i}$ generated by the proposed algorithm for the data group i
- L_{sk} : length (in bits) of the sub-key sk_i of the future meta-key
- M: number of data frames per group

Hence the total number of groups N_g (which is the number of keys) is equal to

$$N_g = round(M * \frac{L_{sb}}{L_{km}}) \quad (1)$$

where M $= \frac{N}{N_f}$ is the overall number of frames. The following algorithm illustrates the different steps involved in the encryption and decryption phases (Fig. 3):

Encryption Phase and Key Exchange (Client Side)

- Use the Symmetrical Metaheuristic ciphering to generate the first metaheuristic key $k_{m,1}$.
- Subdivide $k_{m,1}$ into sub-keys sk_i of the same size $(k_{m,1} = [sk_{1,1}, sk_{2,1}, ..., sk_{M,1}])$.

For j = 1: N_g, If j = 1 (First data group)

For i = 1: M,

- the sub-key $sk_{i,j}$ to the end of the data frame $F_{i,j}$: this results in a composite data frame $[F_{i,j}, sk_{i,j}]$
- Encrypt the composite data frame $[F_i, sk_i]$ using the metaheuristic key k_H obtained in the handshake protocol, this results in an encrypted frame $EF_{i,j} = encrypt([F_{i,j}, sk_{i,j}], k_H)$

End Else For i = 1:M,

- Append the sub-key $sk_{i,j}$ to the end of the data frame $F_{i,j}$: this results in a composite data frame $[F_{i,j}, sk_{i,j}]$
- Encrypt the composite data frame $[F_{i,j}, sk_{i,j}]$ using the metaheuristic key $k_{m,j-1}$, this results in an encrypted frame $EF_{i,j} = encrypt([F_{i,j}, sk_{i,j}], k_{m,j-1})$

End

End

- Use the Symmetrical Metaheuristic ciphering to generate the metaheuristic key $k_{m,j+1}$ for the next data group.
- Subdivide $k_{m,j+1}$ into sub-keys sk_i of the same size $k_{m,j+1} = [sk_1, sk_2, ..., sk_M])$.

End

Decryption Phase and Keys Reproduction (Side Side) For j = 1: N_g, If j = 1 (First data group) For i = 1: M,

- Decrypt the received composite data frame $EF_{i,j}$ using the metaheuristic key k_H obtained in the handshake protocol, this results in an decrypted frame $[F_i, sk_{i,j}] = decrypt([EF_{i,j}, k_H)$
- Decompose the decrypted frame $[F_{i,j}, sk_{i,j}]$ into two parts: $F_{i,j}$ and $sk_{i,j}$

End Concatenate all the frames F_i into one data group. Concatenate all the sub-keys $sk_{i,j}$ into one key km,1 Else For i = 1: M,

- Decrypt the received composite data frame $EF_{i,j}$ using the metaheuristic key $k_{m,j-1}$, this results in an decrypted frame $[F_{i,j}, sk_{i,j}] = decrypt(EF_{i,j}, k_{m,j-1})$
- Decompose the decrypted frame $[F_{i,j}, sk_{i,j}]$ into two parts: $F_{i,j}$ and $sk_{i,j}$

End Concatenate all the frames $F_{i,j}$ into one data group. Concatenate all the sub-keys $sk_{i,j}$ into one key $k_{m,j}$ End

End

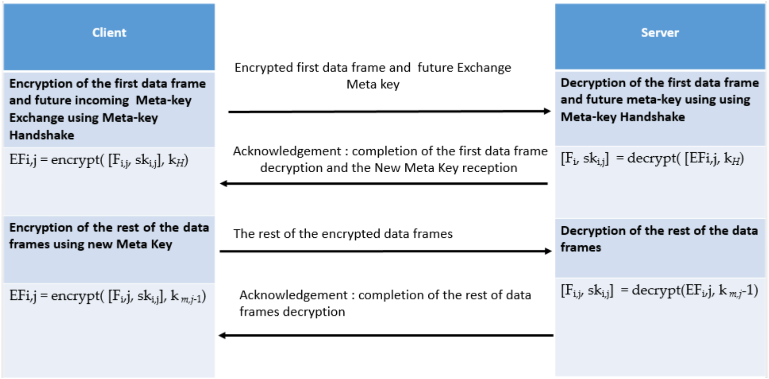

Client
Server
Encrypted first data frame and future Exchange Meta key
Encryption of the first data frame and future incoming Meta-key Exchange using Meta-key Handshake
Decryption of the first data frame and future meta-key using using Meta-key Handshake
EFi,j = encrypt([$F_{i,j}$, $sk_{i,j}$], k_H)
Acknowledgement : completion of the first data frame decryption and the New Meta Key reception
[F_i, $sk_{i,j}$] = decrypt([EFi,j, k_H)
Encryption of the rest of the data frames using new Meta Key
The rest of the encrypted data frames
Decryption of the rest of the data frames
EFi,j = encrypt([F_i,j, $sk_{i,j}$], $k_{m,j}$-1)
Acknowledgement : completion of the rest of data frames decryption
[$F_{i,j}$, $sk_{i,j}$] = decrypt(EF$_i$,j, $k_{m,j}$-1)

Fig. 3. Key exchange and encryption process.

IV. Security Analysis and Performance

The main purpose of this section is to evaluate the robustness of the proposed TLSM protocol.

A. Symmetrical Metaheuristics Ciphering Execution Time

In the following Table I, in detail, the execution time of Symmetrical metaheuristics ciphering compared to 3DES, AES is presented. These results prove that Symmetrical metaheuristics ciphering needs less time to encrypt the different blocks compared to the required time to be encrypted by the other encryption standards.

TABLE I. Symmetrical Metaheuristics Ciphering Execution Time.

Block size (bytes)	AES	3DES	SMC
16020	1.62	2.18	0.69
34280	2.66	2.43	0.65
72620	3.78	6.15	1.34
157440	6.31	10.90	2.52
224116	7.20	12.56	3.93

B. Sha-Sbox Robustness

Sha-sbox is a hash function designed by the combination of SHA-256 and DES [9]. The compression function is based on permutation and substitution (S_BOX). Sha-sbox has a good avalanche effect [10] (i.e. if one bit changed in the initial message input, it affects the half size of the output). Sha-sbox[9] is efficient, fast, resistant to attack and resist to differential and linear cryptanalysis. Its Security level against birthday attack is 2^{128} bits. Furthermore, Sha-sbox[9] uses in its process uses Fps operations that are not costly in execution time. The modular addition applied to the output from permutations and substitution at every round, increases the level of security and avoid collision attack.

C. Meta Secret Key Security Analysis

Symmetrical metaheuristics ciphering [4] generates keys valid for only one session. For this matter, to encrypt a data frame using Symmetrical metaheuristics ciphering in t sessions, k_t keys are required to be generated. Since the robustness of an encryption algorithm is related to the robustness of its secret key, the probability to find the right key P_A is the maximum number of keys K_N by the expression (2):

$$P_A = \frac{1}{K_N} \tag{2}$$

Where K_N denotes the maximum number of different keys generated by the Symmetrical metaheuristics ciphering. Denote by P_A the probability to find the right key using the brute-force attack. Since in Symmetrical metaheuristics, ciphering each data frame is allocated by 8 bits and each data frame of length N samples is encrypted using one meta-secret key. This meta-secret key size is defined by $8N$ bits (Table II).

TABLE II. KEYS SIZE AND BRUTE-FORCE ATTACK COMPLEXITY.

Frame Length	P_A	K_N	Key Size (in bit)	Brute force attack
40	$1.23e-48$	40!	320	2320
80	$1.40e-119$	80!	640	2640
160	$2.12e-285$	160!	1280	21280
320	$4.73e-665$	320!	2560	22560
640	$1.55e-1520$	640!	5120	25120

It is worth to mention that the shortest meta-secret key generated (320 bits) is better than the current AES key 256-bit which is the actual standard for encryption. It is hard to break this meta-secret key using the brute force attack.

D. Key Session Security Analysis

The protection provided by a symmetric encryption algorithm is related to the length of the key that is expressed in bits. In fact, the length of the key quantifies the maximum number of operations needed to find the right key. It is therefore an essential point for the security of the system. The proposed algorithm generates a number of keys in one session that depends on four parameters the size of data exchanged, the size of the processed blocks of data, the sub-blocks size of the future meta-key and the meta-key size. The results in Table III obtained using the general formula (1). Table III shows the number of keys N_g generated for different data lengths (in bits) and different sub-key sk_i of the future meta-key as following. Here a data frame length N_f of 1000 Bytes and a Metaheuristic key $k_{m,i}$ of 320 bits are choosen.

TABLE III. MAXIMUM NUMBER OF KEYS GENERATED BY THE SYMMETRICAL METAHEURISTICS CIPHERING, K_N.

	16020	34280	72620	157440	224116
40	2	4	8	16	25
80	4	7	15	32	45
160	7	14	30	63	90
320	13	28	59	126	180
640	27	55	117	252	359

The metaheuristic encryption system in its initial version has proven its resistance against most known attacks [11]. We have proposed in this paper a new technique for generating and sharing symmetric keys in the data encryption phase. The TLSM record protocol allows generating an additional number of symmetric keys as the data are processed in the encryption phase. This number depends on the size of the processed data and other parameters that are related to the future meta-keys. In Table III a data size of 224116 bytes is combined with 320-bit encryption keys, data frame size of 1000 bytes and 640-bit sub-keys. With these parameters the system can generate up to 359 different meta-keys, a number that is relatively huge. In addition to the basic security of the SMC encryption system, the new proposed technique increases the security of the Protocol record and hence the overall security of the TLSM Protocol while compared to the standard versions of the TLS Protocol.

V. CONCLUSION

We presented in this paper a new protocol termed TLSM to improve the end-to-end secure communication. It uses a previously developed Metaheuristic symmetric ciphering algorithm in order to encrypt data, a recently presented hash function SHA-SBOX and on a new method for private key exchange. The use of these new algorithms in TLS process made the proposed protocol fast and secure as proven by the simulation results. Providing a high performance compared to TLS previous version, this proposed protocol TLSM proved to be robust and not execution-time consuming. In order to continuously improve the TLS protocol and the new TLSM version proposed in this article, it is focused on the cryptanalysis of the encryption suites used in the handshake protocol. In case of vulnerability found or needed improvement, it is suggested to try to design other encryption suites alternatives more robust and performing to ensuring maximum security of data communication.

REFERENCES

[1] Krawczyk, Hugo, Kenneth G. Paterson, and Hoeteck Wee. *On the security of the TLS protocol: A systematic analysis.* Advances in Cryptology–CRYPTO 2013. Springer, Berlin, Heidelberg, 2013. 429-448.

[2] Ronen, E., Gillham, R., Genkin, D., Shamir, A., Wong, D., Yarom, Y. *The 9 Lives of Bleichenbacher's CAT: New Cache ATtacks on TLS Implementations.*2019

[3] Jonsson, Jakob, and Burton S. Kaliski. *On the Security of RSA Encryption in TLS.* Annual International Cryptology Conference. Springer, Berlin, Heidelberg, 2002, p. 127-142

[4] Kaddouri, Zakaria, Mohamed Amine Hyaya, and Mohamed Kaddouri. *A New Cryptosystem using Vigenere and Metaheuristics for RGB Pixel Shuffling.* Coordinates 25.4 2017: 255.

[5] Eronen, Pasi, and Hannes Tschofenig. *Pre-shared key ciphersuites for transport layer security (TLS)*. No. RFC 4279. 2005.

[6] Gupta, Vipul and Wurm, Michael and Zhu, Yu and Millard, Matthew and Fung, Stephen and Gura, Nils and Eberle, Hans and Shantz, Sheueling Chang, *Sizzle: A standards-based end-to-end security architecture for the embedded internet*, 2005, Sun Microsystems, Inc.

[7] Fardan, N., Bernstein, D. J., Paterson, K. G., Poettering, B., Schuldt, J. C. *On the Security of RC4 in TLS.* In Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13). 2013. (pp. 305-320).

[8] Handa, Arun. *System engineering for IMS networks.* Newnes, 2009.

[9] Zakaria Kaddouri, Fouzia Omary, Abdollah Abouchouar And Mohssin Daari *New Compression Function To Sha-256 Based On The Techniques Of Des*, Journal of Theoretical and Applied Information Technology, 2013, Volume 5,pages 230 – 234.

[10] Castro, J. C. H., Sierra, J. M., Seznec, A., Izquierdo, A., Ribagorda, A. *The strict avalanche criterion randomness test.* Mathematics and Computers in Simulation, 2005, 68(1), 1-7.

[11] Paterson, K. G., Ristenpart, T., Shrimpton, T. *Tag size does matter: Attacks and proofs for the TLS record protocol.* In International Conference on the Theory and Application of Cryptology and Information Security. Springer, Berlin, Heidelberg. (2011, December). (pp. 372-389).

[12] Paulson, Lawrence C. *Inductive analysis of the Internet protocol TLS.* ACM Transactions on Information and System Security (TISSEC) 2.3 (1999): 332-351.

[13] Jiao, R., Ouyang, H., Lin, Y., Luo, Y., Li, G., Jiang, Z., Zheng, Q. *A Computation-Efficient Group Key Distribution Protocol Based on a New Secret Sharing Scheme.* Information, (2019). 10(5), 175.

[14] Chen, H. P., Gonzalez, E., Saez, Y., Kish, L. *Cable capacitance attack against the KLJN secure key exchange.* Information, (2015). 6(4), 719-732.

[15] STANEK, Martin. *Secure by default-the case of TLS.* arXiv preprint arXiv:1708.07569, 2017.

[16] AVIRAM, Nimrod, SCHINZEL, Sebastian, SOMOROVSKY, Juraj, et al. *DROWN: Breaking TLS Using SSLv2.* In : 25th USENIX Security Symposium (USENIX Security 16). 2016. p. 689-706.

[17] SHEFFER, Yaron, HOLZ, Ralph, et SAINT-ANDRE, Peter. *Summarizing known attacks on transport layer security (TLS) and datagram TLS (DTLS).* 2015.

Arabic Lexicon Learning to Analyze Sentiment in Microblogs

Mahmoud B. Rokaya[1]
Ahmed S. Ghiduk[2]
Department of Information Technology, Taif University
Taif, KSA

Mahmoud B. Rokaya[3]
Department of Mathematics, Faculty of Science
Tanta University
Tanta, Egypt

Ahmed S. Ghiduk[4]
Department of Mathematics and Computer Science, Faculty of Science, Beni-Suef University
Beni-Suef, Egypt

***Abstract*—The study and classifying of opinions distilled from social media is called sentiment analysis. The goal of this study is to build an adaptive sentiment lexicon for Arabic language. Based on those lexicons the sentiments polarity classification can be improved. The classification problem will be stated as a mathematical programming problem. In this problem, we search a lexicon that optimizes the classification accuracy. A genetic algorithm is presented to solve the optimization problem. A meta-level feature is generated based on the adaptive lexicons provided by the genetic algorithm. The algorithm performance is supported by using it alongside n-gram features and Bing liu's lexicon. In this work, lexicon-based and corpora-based approaches are integrated, and the lexicons are produced from the corpus. Five data sets are tested through experiments. The sentiments in all data sets are classified based on five polarity levels. A better understanding of words sentiment orientation, social media users' culture and Arabic language can be achieved based on the lexicons generated by the proposed algorithm. Since stop words can contribute and add to the sentiment polarity, stop words will be considered and will not deleted. The results show that the F-measure is greater than 80 % in three data sets and the accuracy is greater than 80 % for all data sets. The proposed method out-performs the current methods in the literature in two of the datasets. Finally, in terms of F-measure, the proposed methods achieved better results for three datasets.**

Keywords*—Sentiment analysis; sentiment lexicon; social media; twitter; optimization; mathematical programming; genetic algorithm; evolutionary computation; Arabic language

I. Introduction

The subjects of sentiment analysis are the study of opinions and its related concepts such emotions attitudes evaluations and sentiments. For the first time in humanity history, we have that massive volume of recorded data that reflects the opinions, emotions and attitudes of people around the globe. This came from Twitter, reviews, social networks, forum discussions, blogs and microblogs. So, it is natural that the field of sentiment analysis is emerged.

In business, sentiment analysis addresses the problem of studying the customer opinions regarding products through analyzing and extracting opinion from products reviews. However, most current algorithms which developed for the business purpose are not suitable to analyze sentiments in social domain.

The objective of Sentiment Classification task is to take a piece of text written by an author regarding a topic and determine the author general feeling toward this topic, whether this felling be positive or negative.

The current work tries to improve classification of sentiments in microblogs based on building sentiment lexicons. The sentiment classification problem is written as an optimization problem, finding optimum sentiment lexicon is the goal of the optimization process. The solution will be produced based on proposed genetic algorithm to find lexicons to classify text. Then, extraction of a meta-level feature will be done based on it. The experiments are conducted on several Arabic datasets. A better understanding of the Arabic language and culture of Arab Twitter users and sentiment orientation of words in different contexts can be achieved based on the sentiment lexicons proposed by the algorithm.

Since adaptive lexicons are developed in this work, the trends in the ever-changing environment of Twitter can be captured [1]. Updating the lexicons to adapt with the changes in the culture of the users can be done easily. For example, based only on one feature, the results of the proposed method are promising.

Considering real benefits, to understand the social media and their words context in known domains gives the users the ability to use the words in their messages in more effective transmission methods. Similarly, this idea might be used in producing lexicons for languages that do not own one. In analogues with this, this method can be employed to calculate the sentimental scores for same terms in different contexts and websites. The modification of the method for strength and emotion classification will be explored. Based on the method, it is planned to generate lexicons for the Arabic language.

The rest of this paper is organized as follows: Section II presents the related work; Section III presents the methods including. Experiments, results, discussion is presented in Section IV. Finally, the conclusion and main results are presents in Section V.

II. Related Work

In the proposed method, we try to develop an adaptive lexicon for sentiment analysis; the Statistical methods for sentiment analysis, lexicons-based approaches and evolutionary methods are explored.

Statistical methods have been developed based on the following observation. If two words frequently appear together within the same context, they will have the same polarity. So, by calculating a word relative frequency of co-occurrence with special words for a given word, the polarity of this word can be determined. The performance of these algorithms did not give the same or even near results when applied to training data labeled with emotions which has the potential of being independent of domain, topic and time [2].

In that area, many approaches that address different dimensions of opinions, such as subjectivity, polarity, intensity and emotion were proposed to extract sentiment indicators from natural language texts, whether these indicators are at syntactic or semantic levels. Mohammad and Turney, 2013, conducted experiments on how to formulate the emotion-annotation questions and show that asking if a term is associated with an emotion leads to markedly higher inter annotator agreement than that obtained by asking if a term evokes an emotion [3].

T. Wilson, et al., 2005, presented an approach to phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions [4]. M.M. Bradley and P.J. Lang, 2009, developed a set of verbal materials that had been rated in terms of pleasure, arousal, and dominance to complement the existing International Affective Picture System [5].

Despite that classifying manually will give the most accurate results. It is more than difficult to use manual methods in the labeling process for determining the polarity of comments or posts of users in social media. For this reason, some papers use emoticons as labels [6 and 7]. In [8], the author discussed how this method will produce much noise. Using emoticons, Go et al., 2010, distilled 1600,000 tweets from Twitter dataset [7].

Liu et al., 2012, presented a dataset and used a method of labelling that depends on using emoticons and manual classification [9]. Da Silva et al., 2014, created a classifier ensemble for Twitter sentiment classification [10]. Hu et al., 2013, combined the networked data to benefit from emotional spread in sentiment classification [11]. In [12] features that depend on concepts of semantic are combined with the training set [13]. In (Bravo-Marquez et al., 2013), different approach that employs meta-level features for social media sentiment classification is used, namely for twitter. In this method, different features of words are used for polarity and subjectivity classification. Kaewpitakkun et al., 2014, created a lexicon that finds scores for objective and out of vocabulary words, and used a calculation method that depends on weighting scheme for features [14]. A method that depends on distilling patterns of terms and phrases was developed by Saif et al., 2014, for evaluating those terms and phrases on tweet-level and entity-level sentiment analysis [15]. Feature learning approach was introduced by Baecchi et al., 2015. They used this method for classification of tweets. Namely, they targeted posts that might contain pictures [16]. An unsupervised Learning framework was proposed by Hu et al., 2013. In this method, they combined emotional signals, in Twitter datasets [11]. In [17], a sentiment scoring function was used for classification of tweets. Combination of social connections as well as social emotions between users between posts of the same author was employed by Wu et al., 2016 to get better accuracy [18].

Despite, sentiments are implicitly expressed through patterns, dependencies among words in tweets and latent semantic relations, most existing approaches to Twitter sentiment analysis suppose that sentiment is explicitly expressed through affective words. Also, these methods do not consider that words' sentiment orientations and strengths change continuously throughout various contexts in which the words appear.

Sentiment lexicons can be defined as: those groups of terms and phrases that are assigned numeric scores, which give the sentiment emotional value of a term or phrase. Some lexicons, simply, allocate labels for each term or phrase. These labels are either to be positive or negative. For example, we can report Bing Liu's lexicon as the most known lexicon that uses this simple method. Many studies tried to establish lexicons for sentiment analysis [9, 19, 20, 21, 22 and 23].

Lexicon-based approaches to Twitter sentiment analysis becomes more popular because of their simplicity, domain independence, and good performance. These approaches depend on sentiment lexicons, where a list of words is marked with fixed sentiment polarities; for example, [17, 24, 25, 26 and 27]. Arora et al., 2010, and Govindarajan, 2013, used a hybrid of Naive Bayes classifier and genetic algorithm for classification of movie reviews [28].

For Arabic sentiment analysis, Hossam et al., 2015, presented a sentiment analysis based on two lexicons. The first is a lexicon for adjectives and adjective nouns. The other lexicon contains the known idioms. They developed a method to expand the lexicon from seeds or words and idioms. The method reflects a static lexicon with fixed values for the polarity of each term. Also, they depend heavily on a translated version of HU-LUI lexicon [29]. Haidy et. al., 2017, used a hybrid method to determine the sentiment polarity of a tweet. In the first phase they used a lexicon to classify a set of tweets. The result of this phase is the input of the second phase. The lexicon was composed of two parts. The first is a lexicon for words; the second is a lexicon of idioms [29]. Al-Ayyoub and Essa, 2015, presented a sentiment analysis based on lexicon approach was adopted. The polarity of a given word is got from the corresponding English translation. Stop words are deleted with consideration of some stop words that can affect the polarity of a given word. The lexicon words and sentiment expression are stemmed. Using the polarity of the translated terms will reduce the functionality of the words, also neglecting the stop words, which contribute in the total meaning that the author wants to give [30].

Most of these works stated that they follow a supervised or unsupervised leaning approach without mentioning the training phase and testing phases in their works. To say that lexicon-based approach is an unsupervised approach is not correct in general. In this work, no translation will be applied to get the polarity of words. Also, the proposed method builds a dynamic lexicon where the polarity of the words related to the corps.

The polarity of the same word can be different from corpus to another and can be changed for the same topic by adding more and more sentiments. Also, all these works classified the sentiments into two classes, +ve and –ve classes. In the current work the level of polarity is considered, the sentiment polarity can be strong +ve, +ve, netural, -ve and strongly –ve.

III. The Method

Based on one feature, namely *AAL* (Adaptive Arabic Lexicon), this work tries to find optimized Arabic lexicon. The problem will be written as an optimization problem and the method of optimization will be genetic algorithms. The problem can be stated as: find the lexicon that minimizes the error of polarity classifications for a given set of texts. Suppose the set of lexicons is AL and the set of texts is T. For a given text $t_i$in T and a lexicon $l_j$in $AL$the score of t_i with respect to $l_j$is the sum of the scores of all words in $t_i$with respect to l_j. $AAL_{l_j}(t_i) = \sum_{w \in t_i} S_{l_j}(w)$, $S_{l_j}(w)$ is the score of the word $w$in the lexicon l_j. $t_i$is classified based on the value of $AAL_{l_j}(t_i)$ according to:

$$Pr\ e\ dictedClass(t_i, l_j)$$

$$= \begin{cases} \text{strongly } +\text{ve} & \text{if } \frac{\max_{\text{AAL}}}{2} < AAL(t_i, l_j) \le \max_{\text{AAL}} \\ +\text{ve} & \text{if } \frac{\max_{\text{AAL}}}{4} < AAL(t_i, l_j) \le \frac{\max_{\text{AAL}}}{2} \\ \text{neutral} & \text{if } \frac{\min_{\text{AAL}}}{4} \le AAL(t_i, l_j) \le \frac{\max_{\text{AAL}}}{4} \\ -\text{ve} & \text{if } \frac{\min_{\text{AAL}}}{2} \le AAL(t_i, l_j) < \frac{\min_{\text{AAL}}}{4} \\ \text{strongly -ve} & \text{if } min_{AAL} \le AAL(t_i, l_j) < \frac{\min_{\text{AAL}}}{2} \end{cases}$$

Where

$$max_{AAL} = \underset{t_i \in T}{max} AAL_{l_j}(t_i)$$

And

$$min_{AAL} = \underset{t_i \in T}{min} AAL_{l_j}(t_i)$$

The accuracy $AC_{l_j}(T)$ of a lexicon l_j for the set of texts $T$is ratio of correctly classified texts in T, $NCCT$, to the total number of texts in T,NT:

$$AC_{l_j}(T) = \frac{NCCT}{NT}$$

So, the optimization problem can be written as: find $\mathrm{l_{best}} = \underset{\mathrm{l_j \in AL}}{\mathrm{arg\ max}}\, \mathrm{A\,C_{l_j}(T)}$

Fig. 1 shows how the above classification works. To solve the optimization problem as a genetic optimization problem, we need to define the fitness function; if we used the accuracy function as the fitness function then the algorithm will try to maximize the value of the accuracy function more than improving the classification accuracy. To get a better approach, the concept of punishment and reward will be used. This means that, if the a given text is classified correctly, then the lexicon will be rewarded by adding a positive value to the fitness function and if it did not classify the text correctly, the lexicon will be punished by adding a negative value to the fitness function. Let the fitness function be $FAAL_{AL}(T)$. The increment function $INC_{l_j}(t_i)$ is given by:

$$INC_{l_j}(t_i)$$

$$= \begin{cases} |AAL(t_i, l_j)| & \text{if } l_j \text{ correctly classified } t_i \text{ and } |AAL(t_i, l_j)| \neq 0 \\ 1 & \text{if } l_j \text{ correctly classified } t_i \text{ and } |AAL(t_i, l_j)| = 0 \\ -|AAL(t_i, l_j)| & \text{if } l_j \text{ incorrectly classified } t_i \text{ and } |AAL(t_i, l_j)| \neq 0 \\ -1 & \text{if } l_j \text{ incorrectly classified } t_i \text{ and } |AAL(t_i, l_j)| = 0 \end{cases}$$

The fitness function $FAAL_{AL}(T)$ is given by:

$$FAAL_{l_j}(T) = \sum_{t_i \in T, l_j} INC_{l_j}(t_i),$$

where$AL_G$is the chromosomes of the current generation

Fig. 2 shows an example of the classification of a sentiment based on a given lexicon. In this example the used sentiment is "أردوغان: تعرضنا لمحاولة اغتيال اقتصادي في أغسطس" ("Erdogan: We were hit by an economic assassination attempt in August"). The algorithm distills the polarity of each term from the lexicon, add all values then it classifies the sentiment based on the proportional place of this value between min AAL and max AAL.

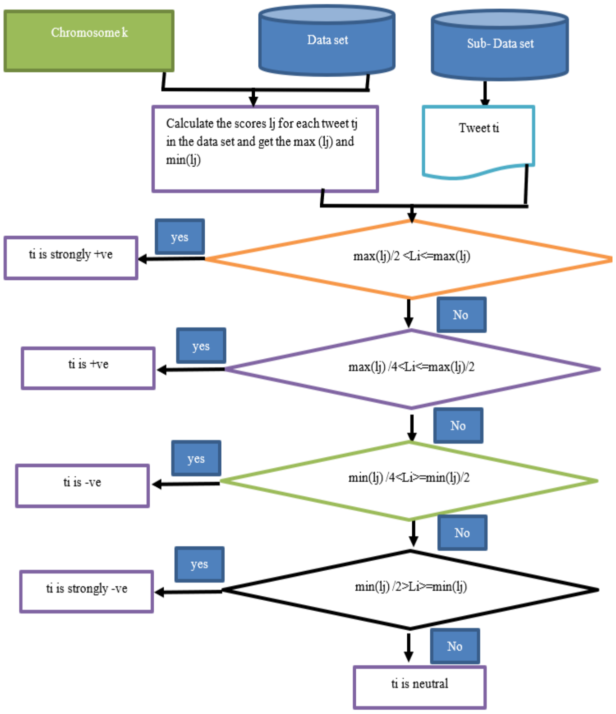


Fig. 1. How to Classify a Given Sentiment.

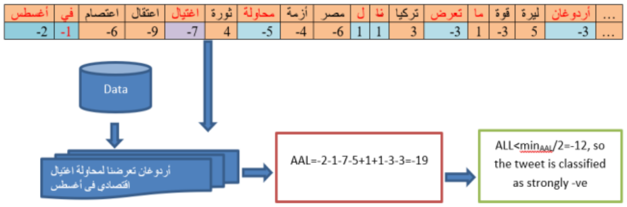


Fig. 2. Example of Classifying a Given Sentiment

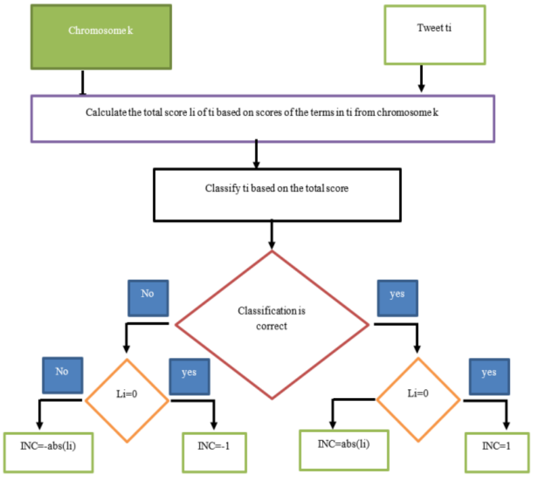

Chromosome k
Tweet ti
Calculate the total score li of ti based on scores of the terms in ti from chromosome k
Classify ti based on the total score
Classification is correct
No
yes
No
yes
Li=0
Li=0
yes
INC=-abs(li)
INC=-1
INC=abs(li)
INC=1

Fig. 3. Calculating INC.

Fig. 3 illustrates how to calculate INC. The following algorithm explains how to calculate the $FAAL_l$ (T) function of chromosome $l$in data set T

Algorithm 1 Fitness function of chromosome $l$in data set l

```
Fitness($l$,$T$)
$f$ =0
for each $t_i$in $T$
    $AAL = 0$
    for each word $w$in $t_i$
        $AAL = ALL + S_l(w)$ //the score of $w$in chromosome $l$
    end for
    if the value $AAL$makes $t_i$to be classified correctly and $|AAL(t_i, l_j)| \neq 0$
        Then $f = f + |AAL(t_i, l_j)|$
    if the value of $AAL$ makes $t_i$ to be classified correctly and $|AAL(t_i, l_j)| = 0$
        Then $f = f + 1$
    if the value of $AAL$ makes $t_i$ to be classified incorrectly and $|AAL(t_i, l_j)| \neq 0$
        Then $f = f - |AAL(t_i, l_j)|$
    if the value of $AAL$ makes $t_i$ to be classified incorrectly and $|AAL(t_i, l_j)| = 0$
        Then $f = f - 1$
    end if
end for
return $f$
```

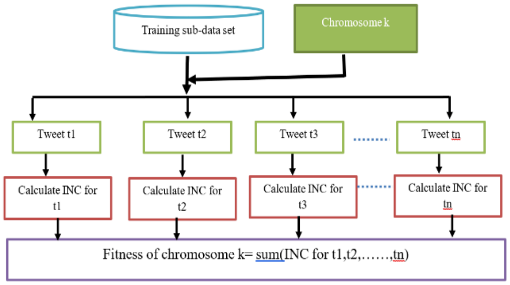


Fig. 4. Calculation of Fitness Function.

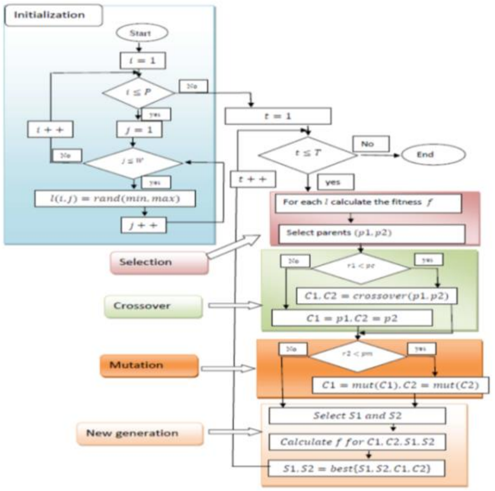

Initialization
Start
$i = 1$
No
$i \le P$
yes
$i++$
$j = 1$
No
$j \le W$
yes
$l(i,j) = rand(min, max)$
$j++$
$t = 1$
$t \le T$
No
End
$t++$
yes
For each l calculate the fitness f
Select parents $(p1, p2)$
Selection
No
$r1 < pc$
yes
$C1, C2 = crossover(p1, p2)$
Crossover
$C1 = p1, C2 = p2$
Mutation
No
$r2 < pm$
yes
$C1 = mut(C1), C2 = mut(C2)$
Select $S1$ and $S2$
New generation
Calculate f for $C1, C2, S1, S2$
$S1, S2 = best(S1, S2, C1, C2)$

Fig. 5. Genetic Algorithm Flowchart.

Fig. 5 shows the details of the genetic algorithm. The genetic algorithm consists of five main parts. The first part is the initialization part where a random population is chosen. The algorithm will choose random vectors, the length of each vector is equal to the number of the unrepeated words and the values are distributed randomly over the interval (minv, maxv). The algorithm checked many values or minv and maxv during the training phase and kept the values that gave the best results. The second phase calculates the fitness value for each chromosome and chooses the next generation. The last phase includes crossover, mutation and replacement to generate the new generation. Based on the roulette wheel strategy, lexicon with higher fitness values are more likely to be selected. The crossover is implemented randomly. If a selected random number between 0 and 1 is less than a given probability value, then a crossover for the current parents will produce the next children otherwise the new children will be identical to their parents. A mutation is implemented, for a random selected value between 0 and 1; the mutation for the resulting children will be applied if the number is less than a specific probability. Finally, a replacement will be applied; Lexicons with lower fitness values are more likely to be replaced. Fig. 4 gives the details of the calculation of the fitness function.

IV. Experiments

In this section, data sets, parameters and results are introduced. The results of using the proposed method on the datasets are analyzed and reported.

A. Data Sets

AAL was run on five different data sets from tweets of usedrs in Twitter. These data sets were given names, TLC, MBH, NSC, SIE and TRE.

- TLC corpuscontains 701 tweets about the crises of Turkish Lira, 200 +ve , 140 –ve, 130 strongly +v, 131 strongly - ve and 100 neutral tweets.
- MBH contains 1073 tweets about Muslims brotherhood. There was 325 strongly +ve, 311 strongly –ve, 168 +ve, 131 –ve and 138 neutral tweets.

- NSC corpus contains 613 tweets about New High School regulations in Egypt. 121 strongly +ve, 115 strongly +ve, 175 +ve, 111 –ve and 91 neutral tweets.
- SIE contains 608 tweets about the last Egyptian elections. 81 strongly +ve, 211 strongly –ve. 25 +ve, 240 –ve and 51 neutral tweets.
- TRE consists of 982 tweets about the American elections. 95 strongly +ve, 315 strongly –ve, 74 +ve, 357 –ve and 141 neutral tweets. Table I summaries the data sets information.

Fig. 8 shows how the program is running. A program was written to implement the proposed algorithm. A k-fold method was used for the algorithm with k=15. Each time the data sets are divided into 15 subsets and 14 of these subsets were used as the training set, the 15th subset was used as the validation set. This process was repeated 15 times, each time one subset was used as a validation set and the remaining 14 sets were used as the training set. The final result is the average of the 15 running's of the algorithm. The range of terms polarity, crossover range and mutation rate were set as follows:

Fig. 6 shows how the crossover process is applied. Some cells are chosen randomly from each chromosome. The chosen cells from Parent A are replaced by the corresponding cells from Parent B cells in

Fig. 7 shows an example of mutation:

- The range of polarity for each term in the lexicon was set to be between -10 and +10
- A uniform crossover was applied with rate 0.8
- The mutation rate was set to 0.05

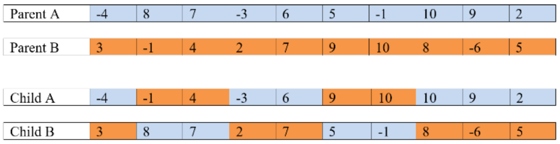
The algorithm was run till no improvement can be achieved. Sets of parameters were chosen to run the algorithm on different data sets. Namely, there were two sets of parameters which were used with two different sets of data. The original sets of data were randomly divided into two equal data sets. Equal here means that the number of sentences in each set is equal to the number of sentences in the other set.

B. Results

In this section, we will provide the results of our approach to build adaptative lexicon in terms of F1-measure for our data sets.

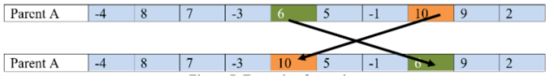
TABLE. I. POSITIVE, NEGATIVE AND NEUTRAL TOTAL NUMBER OF TWEETS IN EACH DATASET

Polarity	DATASET				
	TLC	***MBH***	***NSC***	***SIE***	***TRE***
Strongly Positive	131	325	121	81	95
Positive	200	168	175	25	74
Neutral	100	138	91	51	141
Negative	140	131	111	240	357
Strongly Negative	131	311	115	211	315
Total	702	1073	613	608	982



Parent A	-4	8	7	-3	6	5	-1	10	9	2
Parent B	3	-1	4	2	7	9	10	8	-6	5
Child A	-4	-1	4	-3	6	9	10	10	9	2
Child B	3	8	7	2	7	5	-1	8	-6	5

Fig. 6. Example of Crossover.



Parent A	-4	8	7	-3	6	5	-1	10	9	2
Parent A	-4	8	7	-3	10	5	-1	6	9	2

Fig. 7. Example of Mutation.

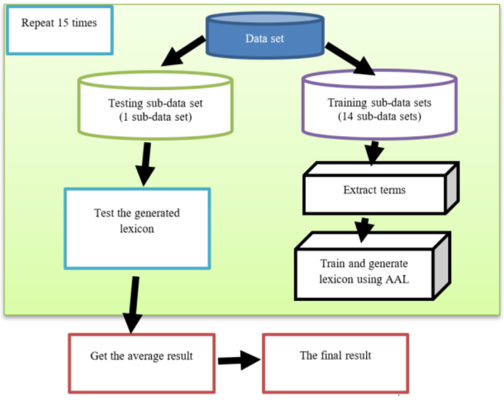


Fig. 8. K-fold Method to Test the Proposed Method.

Table II shows the results of F1-measure and Accuracy values for different mutation and crossover rates on the SIE and TRE datasets. In each case, the best values of crossover and mutation rates were reported.

TABLE. II. THE F1-MEASURE AND ACCURACY VALUES FOR DIFFERENT MUTATION AND CROSSOVER RATES ON THE SIE AND TRE DATASETS

		SIE			
pc	***pm***	***Accuracy***	***F1-Score***	***Accuracy***	***F1-Score***
0.6	0.01	75.51	78.21	81.9	83.92
0.6	0.02	81.98	78.22	80.48	80.38
0.6	0.05	80.83	78.42	84.25	80.74
0.6	0.1	76.34	76.76	83.12	77.9
0.7	0.01	74.65	78.75	82.81	78.53
0.7	0.02	78.69	75.67	81.63	82.08
0.7	0.05	80.13	79.87	83.37	81.86
0.7	0.1	83.44	78.96	82.89	76.79
0.8	0.01	75.99	79.01	80.37	82.39
0.8	0.02	74.87	75.25	77.72	80.72
0.8	0.05	77.81	75.81	81.47	82.02
0.8	0.1	80.67	76.8	81.03	82.13
0.9	0.01	81.38	78.84	81.38	81.24
0.9	0.02	77.54	75.36	79.38	78.07
0.9	0.05	77.26	77.89	81.31	81.16
0.9	0.1	79.81	77.26	83.24	82.77
1	0.01	79.4	80.31	79.91	81.14
1	0.02	80.42	75.51	82.79	75.42
1	0.05	78.83	82.38	81.47	79.62
1	0.1	75.48	77.22	77.66	78.65

For testing mutation and crossover rate settings, we examined different values. For these two datasets Fig. 9 and Fig. 10 show the relation between different parameter values and F-measure. For each dataset and setting, the algorithm was run. The results were reported based on averaging running. From the results we can conclude that the best performance was at values between 0.6 and 0.9 for crossover and at values between 0.05 and 0.1 for mutation. To insure the results independence from crossover and mutation rates, crossover and mutation rates were fixed at 0.8 and 0.06. Reviewing the results in Table III, the proposed method gave good results that outperform the current available methods in many cases. Regarding the number of iterations, a limited number of iterations, 100,000 iterations were enough, and conversion was achieved for small data sets. For big data sets, the conversion was achieved with iterations numbers around 250000 iterations. This leads us to consider iterations number 250000 for all data sets.

TABLE. III. AAL Running Results on All Data Sets

Method	***Accuracy (%)***	***Precision (%)***	***Recall (%)***	***F1 (%)***	***Precision (%)***	***Recall (%)***	***F1 (%)***	***Average F1 (%)***
TLC-dataset								
Bing=Liu-Lexicon	66.0	72.0	27.9	40.2	66.3	90.6	76.6	58.4
Random-search	51.8	39.3	42.6	40.9	65.9	55.3	60.2	50.5
AAL	83.1	77.1	79.3	78.2	85.5	84.2	84.9	81.5
AAL-SW	79.4	74.6	74.6	74.6	85.0	84.6	84.8	79.7
AAL+1,2,3-grams	86.5	81.9	78.3	80.1	88.7	88.8	88.8	84.4
AAL+lex	84.7	84.1	79.0	81.5	91.5	87.3	89.4	85.4
AAL+lex+1,2,3-grams	86.5	86.1	81.4	83.7	88.7	88.9	88.8	86.2
Best-reported-results-from-the-literature	80.7	75.2	77.6	76.4	85.5	88.9	87.2	81.8
MBL-dataset								
Bing=Liu-Lexicon	69.5	73.8	73.1	73.4	73.8	71.3	72.6	73.0
Random-search	52.9	45.2	39.7	42.3	52.6	60.8	56.4	49.4
AAL	83.1	77.7	88.7	82.8	86.8	78.1	82.2	82.5
AAL-SW	81.1	80.4	84.2	82.3	77.8	79.8	78.8	80.5
AAL+1,2,3-grams	85.3	83.6	87.2	85.3	87.6	84.6	86.1	85.7
AAL+lex	88.0	82.2	85.7	83.9	86.0	89.1	87.5	85.7
AAL+lex+1,2,3-grams	85.9	86.4	89.4	87.9	89.9	88.8	89.3	88.6
Best-reported-results-from-the-literature	97.6	83.9	86.8	85.3	87.1	85.6	86.3	85.8
NSC-dataset								
Bing=Liu-Lexicon	81.6	22.9	33.5	27.2	60.8	83.2	70.3	48.7
Random-search	59.1	24.8	40.0	30.6	60.8	44.4	51.3	41.0
AAL	75.0	39.3	33.6	36.2	65.0	65.7	65.3	50.8
AAL-SW	72.1	39.3	30.8	34.6	59.0	66.4	62.5	48.5
AAL+1,2,3-grams	82.9	61.3	37.4	46.5	63.9	72.5	67.9	57.2
AAL+lex	80.5	42.1	35.5	38.5	64.0	65.0	64.5	51.5
AAL+lex+1,2,3-grams	82.5	62.8	40.9	49.5	66.3	76.5	71.0	60.3
Bes-reported-results-from-the-literature	80.4	55.8	60.4	58.0	65.9	69.7	67.7	62.9
SIE-dataset								
Bing=Liu-Lexicon	69.5	61.2	23.7	34.1	66.4	94.8	78.1	56.1
Random-search	52.7	62.6	61.5	62.1	36.0	40.2	38.0	50.0
AAL	78.9	76.4	71.3	73.8	85.6	87.5	86.5	80.1
AAL-SW	77.5	68.4	65.0	66.7	78.2	82.9	80.5	73.6
AAL+1,2,3-grams	80.0	75.0	71.7	73.3	83.0	86.2	84.6	79.0
AAL+lex	80.4	77.3	68.5	72.6	80.6	84.5	82.5	77.6
AAL+lex+1,2,3-grams	83.7	78.2	74.4	76.2	87.1	90.0	88.5	82.4
Best-reported-results-from-the-literature	82.9	75.9	67.4	71.4	82.8	87.6	85.2	78.3
TRE-dataset								
Bing=Liu-Lexicon	72.8	68.2	91.2	78.1	84.9	57.7	68.7	73.4
Random-search	53.8	55.1	47.6	51.1	49.6	53.3	51.3	51.2
AAL	77.8	78.9	76.2	77.5	79.1	78.0	78.5	78.0
AAL-SW	79.3	82.6	75.0	78.6	78.8	82.8	80.7	79.7
AAL+1,2,3-grams	82.1	80.6	88.6	84.4	86.2	78.7	82.2	83.3
AAL+lex	78.4	83.3	80.8	82.0	81.3	82.1	81.7	81.9
AAL+lex+1,2,3-grams	85.9	83.2	90.7	86.8	88.5	85.6	87.0	86.9
Best-reported-result-from-the-literature	88.0	85.9	91.8	88.8	84.6	89.2	86.8	87.8

TABLE. IV. ACCURACY AND F1 VALUES FOR 0.95 CONFIDENCE INTERVAL FOR ON THE FIVE DATASETS

	TLC	*MBH*	*NSC*	*SIE*	*TRE*
Accuracy-of-AAL	82.26±2.13	82.81±2.04	67.67±2.2	80.81±1.75	69.91±2.64
F1-Score-of-AAL	79.85±2.42	80.73±2.06	68.65±1.68	75.82±2.18	71.95±2.35
Accuracy-of-AAL+lex+n-grams	87.92±2.17	87.26±2.62	64.74±2.2	82.29±1.69	76.04±2.28
F1-Score-of-AAL+lex+n-grams	84.13±2.27	87.66±3.02	69.95±2.64	81.29±1.88	79.81±2.37

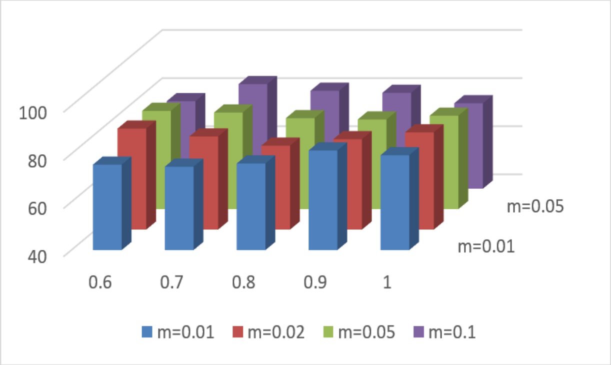


Fig. 9. Values of F1-Measure for Multiple Mutation and Crossover Rates on the SIE Dataset.

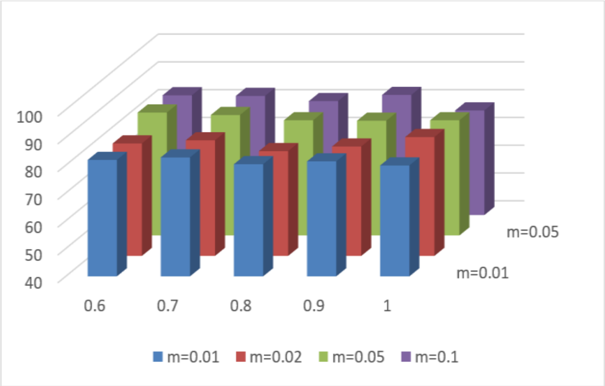


Fig. 10. Values of F1-Measure for Multiple Mutation and Crossover Rates on the TRE Dataset.

C. Discussion

Random search approach and Bing Liu's lexicons are considered the best methods. So, it was natural to compare the performance of the proposed method with these approaches. Table IV shows the comparison results. Best values are bolded. In the random search, based on the representation in the proposed algorithm, a random value is given to initiate a single chromosome. For 250,000 iterations, a neighbor of chromosome is given through changing a single cell in it randomly. If the fitness value of the generated neighbor is higher (based on AAL calculations), the neighbor replaces the original one. A confidence interval is reported since the algorithm is run fifteen times for each fold and it each fold and we have 15 folds. The 0.95 confidence intervals are shown in Table IV. Many variations enhance the AAL performance. AAL-SW is the AAL after removing the stop words. AAL+1,2,3-grams are variations of AAL are the result of applying AAL supported by n-grams features. Enhancing AAL by considering features of meta-level Bing Liu's lexicon produces a modified version of AAL, AAL+lex. Adding n-grams features and metalevel features of Bing Liu lexicon improves the results and makes them better in many measures in the datasets. From Table III, we note that AAL alone could to outperform the other methods in MBH data set. This is due to the clearness of positivity and negativity levels in this data set. However, the worst results of AAL were in NSC data set, also this due to that the level of polarity ambiguity in this data set is the highest among the other data sets. The results reflect a promising result based on using AAL alone. As a classifier, AAL results outperform other classifiers, see [7], [9]. Falsely results in AAL can be explained because of tone of tweet problem. The terms that have low frequency tend to have higher variance when running the algorithm multiple times. Consequently, those terms tend to have improper values. The standard deviation of scores values of sentiment of terms is shown in Table IV.

V. CONCLUSION

In this work, we proposed a genetic algorithm to build an adaptive Arabic lexicon for sentiment analysis. We can report that the F-measure of AAL is 4.13 percentage points better than the average of reported results on the MBH dataset, 3.28 on the TLC dataset, 2.14 on the SIE dataset, and 1.56 on the TRE dataset. AAL achieved accuracy levels better than traditional methods on three data had better accuracy results than state-of-the-art methods on three datasets. For F-measure results, the proposed method achieved better results in four datasets. This work shows that adaptive lexicons can be applied for Arabic language. In fact, the independence of the method from the language is approved. The proposed method can enable better understanding of sentiment words. Since, we did not remove stop words, then this show that all words in Arabic can be considered as sentiment words. In this paper, we approved that writing generating adaptive lexicon as optimization search and applying genetic algorithms to get optimal solution can give an excellent result when applied to Arabic language. It is shown that, AAL can give a high accuracy with small data sets. From the business point of view, the companies can use AAL to create lexicons to help in finding and exploring what users think about. Companies can also use AAL to enrich the knowledge about individual words and their importance; this will increase the effectiveness of manual analysis of sentiments. For example, A supermarket manager can use AAL to create a lexicon for the products and use it for sentiment analysis of their customers behaviors. In this paper, AAL used to analyze the strength of opinions of sentiments. In the future, building a deep net that can apply AAL online with active learning to provide real time adaptive lexicons will be explored.

ACKNOWLEDGMENT

The authors would like to thank Taif University for funding this research under project number: 5607-438-1.

References

[1] H. Keshavarz, M. S. Abadeh, "ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs", Knowledge-Based Systems 122, pp. 1–16, 2017.

[2] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification", in: Proceedings of the ACL Student Research Workshop, ACLstudent'05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 43–48, 2005.

[3] S.M. Mohammad, P.D. Turney, "Crowdsourcing a word–emotion association lexicon", Comput. Intell. 29 (3),pp. 436–465, 2013.

[4] W. J. Wiebe, P. Hoffmann, "Recognizing contextual polarity in phraselevel sentiment analysis", in: Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, Canada, pp. 347–354, 2005.

[5] M.M. Bradley, P.J. Lang, "Affective Norms for English Words (ANEW) Instruction Manual and Affective Ratings", Technical Report C-1, The Center for Research in Psychophysiology University of Florida, 2009.

[6] C. L. Sarmento, M.J. Silva, E. de Oliveira, "Clues for detecting irony in user-generated contents: oh...!! it's so easy;-)", in: Proceeding of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion, Hong Kong, China, pp. 53–56, 2009.

[7] Go, R. Bhayani, L. Huang, "Twitter Sentiment Classification using Distant Supervision", Technical report Stanford University, 2010.

[8] B. Marquez, M. Mendoza, B. Poblete, "Meta-level sentiment models for big social data analysis", Knowl. Based Syst. 69,pp. 86–99, 2014.

[9] Liu, W. Li, M. Guo, "Emoticon smoothed language models for Twitter sentiment analysis", in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Canada, 2012.

[10] N.F.F. da Silva, E.R. Hruschka, E.R. Hruschka, "Tweet sentiment analysis with classifier ensembles", Decis. Supp. Syst. 66, pp. 170–179, 2014.

[11] X. Hu, L. Tang, J. Tang, H. Liu, "Exploiting social relations for sentiment analysis in microblogging", in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 537–546, 2013.

[12] H. Saif, Y. He, H. Alani, "Semantic sentiment analysis of twitter", in: Proceedings of the 11th International Conference on The Semantic Web, ISWC'12, Springer-Verlag, pp. 508–524, 2012.

[13] B. Marquez, M. Mendoza, B. Poblete, "Combining strengths, emotions and polarities for boosting twitter sentiment analysis", in: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, 2013.

[14] Y. Kaewpitakkun, K. Shirai, M. Mohd, "Sentiment lexicon interpolation and polarity estimation of objective and out-of-vocabulary words to improve sentiment classification on microblogging", in: Proceedings of 28th Pacific Asia Conference on Language, Information and Computation, pp. 204–213, 2014.

[15] H. Saif, M. Fernandez, Y. He, H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of Twitter", in: Proceedings of the 9th language resources and evaluation conference (LREC),pp. 810–817, 2014.

[16] B. T. Uricchio, M. Bertini, A. Del Bimbo, "A multimodal feature learning approach for sentiment analysis of social network multimedia, Multimed". Tools Appl, pp. 1–19, 2015.

[17] B. P. Arora, S. Madhappan, N. Kapre, M. Singh, V. Varma, "Mining sentiments from tweets", in: Proceedings of the WASSA, 2012.

[18] F. Wu, Y. Huang, Y. Song, "Structured microblog sentiment classification via social context regularization", Neurocomputing, 175, pp. 599–609, 2016.

[19] L. W. Li, M. Guo, "Emoticon smoothed language models for Twitter sentiment analysis", in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Canada, 2012.

[20] F. Nielsen, "A new anew: evaluation of a word list for sentiment analysis in microblogs", in: Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages, Heraklion, Crete, Greece, 2011.

[21] E. F. Sebastiani, "Sentiwordnet: a publicly available lexical resource for opinion mining", in: Proceedings of the 5th Conference on Language Resources and Evaluation, pp. 417–422, 2006.

[22] B. A. Esuli, F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining", in: Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valletta, Malta, pp. 2200–2204, 2010.

[23] M. Thelwall, K. Buckley, G. Paltoglou, "Sentiment strength detection for the social web", J. Am. Soc. Inf. Sci. Technol, 63 (1),pp. 163–173, 2012.

[24] A. H. Chen, A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums", ACM Trans. Inf. Syst, 26 (3), pp. 12-34, 2008.

[25] B. Gómez, N. Luis Mingueza, M.C. García del Pozo, OpinAIS: "An artificial immune system-based framework for opinion mining", Int. J. Artif. Intell. Interact. Multimed. 3, pp. 25-29, 2015.

[26] A. E. Mayfield, C. Penstein-Rosé, E. Nyberg, "Sentiment classification using automatically extracted subgraph features", in: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, 2010.

[27] Govindarajan, "Sentiment analysis of movie reviews using hybrid method of Naive Bayes and genetic algorithm", IJACR, 3 (4), pp. 139–145, 2013.

[28] H. S. Ibrahim, Sherif M. Abdou and Mervat Gheith, "Sentiment Analysis for Modern Standard Arabic and Colloquial", International Journal on Natural Language Computing (IJNLC), 4(2), pp. 95-109, April 2015.

[29] H. H. Mustafa, A. Mohamed, and D. S. Elzanfaly, "An Enhanced Approach for Arabic Sentiment Analysis", International Journal of Artificial Intelligence and Applications (IJAIA), 8(5), pp. 1-14, September 2017.

[30] M. Al-Ayyoub, S. Bani Essa; I. Alsmadi, "Lexicon-based sentiment analysis of Arabic tweets", International Journal of Social Network Mining (IJSNM), 2(2), pp. 1-14, 2015.

An Efficient Segmentation of Retinal Blood Vessels using Singular Value Decomposition and Morphological Operator

N. C.Santosh Kumar[1], Y.Radhika[2]
Department of Computer Science and Engineering, GIT, GITAM (Deemed to be University), Visakhapatnam, India

***Abstract*—The extensive study on retinal fundus images has become an essential part in medical domain to detect pathologies including diabetic retinopathy, cataract, glaucoma, macular degeneration,etc.which are the major causes of blindness. Automatic extraction of tree-shaped and unique retinal vascular structure from retinal fundus images is most exigent task and when achieved successfully, becomes a perfect tool helping ophthalmologists to follow appropriate diagnostic measures. In this work, a novel scheme to segment retinal tree-like vascular structure from retinal images is proposed using the Singular Value Decomposition's left singular vector matrix of the weighted l*a*b* color model of the input image. The left singular vector matrix which captures the relevant and useful features helps in effective conversion of the input RGB image to gray image. Next, the converted gray image is contrast enhanced using CLAHE method which enhances the tree-shaped vasculature of the retinal blood vessel structure giving a rich contrast gray image. Further processing is carried out normalizing the contrast enhanced gray image by removing the image's background using a mean filter by which blood vessels become brighter. Later, the result of the difference between gray image and normalized filtered image is keyed-in as a constraint to perform ISODATA thresholding which globally segments the foreground vasculature from the image's background then followed by conversion of the resultant image into binary image upon which morphological opened operation is applied to take away small and falsely segmented portions producing accurate segmentation. This new technique got tested upon images contained in DRIVE and STARE databases and a performance metric called "area covered" is also calculated in addition with common metrics for sampled input image. This novel approach is empirically proven and has attaineda segmentation accuracy of 97.48%.**

Keywords*—Singular value decomposition; left singular vector matrix; feature extraction; average filter; ISODAT Athresholding; morphological operators; stare anddrive databases

I. Introduction

Diabetes is such a disease that is affecting much and very rapidly deterioratinghuman's health, globally. Despite the advancements in the field of medicine, this disease will surely get amplified more and would ruin the happiness of the people irrespective of their age, sex, creed, etc. in the future. In automation of the discovery of blood vessels for a perfect analysis, the ideal blood vessels extraction is a complicated job as the appearance of the vessels takes indifferent diameter and width [1]. Thesequantifiable parameters of the varying vessel structure provide a clear specification of the kind of the ophthalmologic disease and its strength. According to [2], a huge number of patientshave lost their eye sight just because of the reach of diabetes on their retina. This spread of such disease is known formally as 'Diabetic Retinopathy'which isdetected by finding the discrepancies in the arrangement of retinal vasculature of retinal images. Most of the medical practitioners have suggested that early discovery and proper diagnosis of such irregularities that exists in blood vessels of retinal imagescouldlessenthe patients'suffering from vision failure. Blood vessels segmentation serves as a source for proper diagnosis of various ophthalmologic and cardiovasculardisorders that includes diabetic retinopathy, glaucoma, hypertension, etc. Vessel segmentation follows either supervised or unsupervised approach and is the primary footstep in performing computer aided diagnosis system development [3] that determines an ophthalmic disease. The outline, manifestation, and the direction of the tree-shaped retinal vasculature make the automation of segmentation a very challenging task. The issues relating to the contrast of the blood vessel structure and its background of the image, the noise presencein the image, and the anomalous composites upon retina like exudates, and lesions, microaneurysms, etc. makes the segmentation a tough task. In this effort, the proposed approach for segmentation ofblood vessels has workedupon varied sized retinal fundus color images and has been proven very efficient in performing the desired segmentation task. Thisunsupervised segmentation method has its roots from a mathematical technique which is well known as 'Singular Value Decomposition (SVD)'and has been exceptional in efficient extraction of blood vessels. These blood vessel features are captured in SVD's left singular vector matrix. Using these extracted features, the image preprocessing is carried out in which transforming the input RGB into grayscale image is done effectively. To overcome the major difficulties in working on green channel of the input retinal image, this successful attemptof using matrix algebra and vector calculus based Singular Value Decomposition is proven to be an outstanding one in image preprocessing phasewhich later makes life easier to moveontopost processing of the image using ISODATAthresholding and Morphological Operation producing an average segmentation accuracy of 97.48%.

The remaining part of this paper is systematically organized with explanation of the proposed work in Section 2. The proposed method's empirical results are furnished in Section 3 with corresponding discussions. Section 4 gives the conclusion of the paper.

II. Materials and Methods

A non-invasive retinal fundus image capturing procedure involves penetration of red, green, and blue lights by the fundus camera into the human eye that eventually get absorbed back by the lens pigments differently. Among the three colors, the green channel in the RGB color image represents a rich contrast in image's blood possessing elements i.e. vessel structure than the background of the image. Most of the research which has been done had used the third portion of the input image i.e. green channel because it's a best and the finest way to exhibit vessels or background contrast. Also, it has the most light than other bands to which human eye is very much sensitive as compared to the red channel, which does possess the brightest color and have low contrast. Other side, blue channel suffers and bear deprived dynamic range [4].But, sometimes green channel also creates hidden noise in the image due to variations in the intensities that infers blurriness in the image and in turn become a challenging task to overcome.The other important reason is that the blood vessels possess very meager local contrast and to overcome this problem in processing an image, various imaging filters have been applied so far.To have an alternative and a replacement to retinal image's green channel, a new procedure has been devised for proper transformationto gray imagefromRGB color image with the help of matrix algebra and vector calculus related mathematical technique called "Singular Value Decomposition (SVD)" which is computed upon input image and later weighted with well-known l*a*b color model. In this paper, left singular vector matrix which is one among the decomposed matrices of the computed Singular Value Decomposition (SVD) (*where SVD takes in l*a*b* processed input image as input*) using the built-in function in MATLAB, captures the required features of blood vessels and arranges them in the decreasing order as the column vectors based on the light intensities provided by the l*a*b* while keeping the non-blood vessels as zeroes simultaneously. This separation and arrangement of blood as well non-blood vessels is one of the magnificent task performed by decomposed left singular vector matrix of SVD.

The sequence of steps depicted in the flowchart is shown in Fig. 1. It gives the step-by-step tasks performed in blood vessels extraction. After reading the input image, it is modeled to l*a*b* color model upon which Singular Value Decomposition (SVD) is computed to capture the required feature set of blood vessels as the ordered column vectors in the decomposed left singular vector matrix. Such feature set which is held by the left singular vector matrix is then contrast enhanced using CLAHE. Later, a mean filter is been used to exclude image's background from the contrast enhanced image. Then, the difference between the contrast enhanced image and filtered image is computed to be used as a threshold level parameter for the ISODATAthresholding which automatically finds a good threshold level to help binarize the whole gray image for segmenting the foreground tree-shaped blood vessel structure. Finally, an open morphological operation removes the smaller sized and unwanted vessels producing a perfect and required segmentation result.

Proposed method's algorithm steps:

The following are sequence of steps of the proposed procedure for segmentation of blood vessels:

Step 1 Read input image
Step 2 Using Singular Value Decomposition convert RGB to gray image
Step 3 Apply CLAHE for contrast enhancement of the gray image ofStep (2)
Step 4 Exclude background of the image by applying Averaging Filter with 9x9 mask.
Step 5 Compute the difference image of Step (3) and Step (4) images to be used as a parameter for ISODATA thresholding which is computed in the Step (6).
Step 6 Find the threshold for the image obtained in Step (5) using ISODATA thresholding Method
Step 7 Convert the resultant image of Step (5) to binary image using the computed threshold of Step (6).
Step 8 Remove small pixels using morphological open operator on the resultant image of Step (7).
Step 9 Realize the segmented image

A. Singular Value Decomposition (SVD)

Any real and a square or a rectangular matrix D in linear algebra, could be decomposed perfectly as a product of three distinct matrices D = PQR^T , where matrices P and R are termed as orthogonal matrices i.e. $P^T P$ = I, $R^T R$ = I and the singular matrix is Q = diagonal (q1, q2 ...). Here, the diagonal entries are D's singular values and each column in P is the D's left singular vectors, and each column in R is the D's right singular vectors. This helpful breakdown of a single matrix into three is called as Singular Value Decomposition (SVD) [5]. Generally, this SVD can be used in three ways in any real world application.

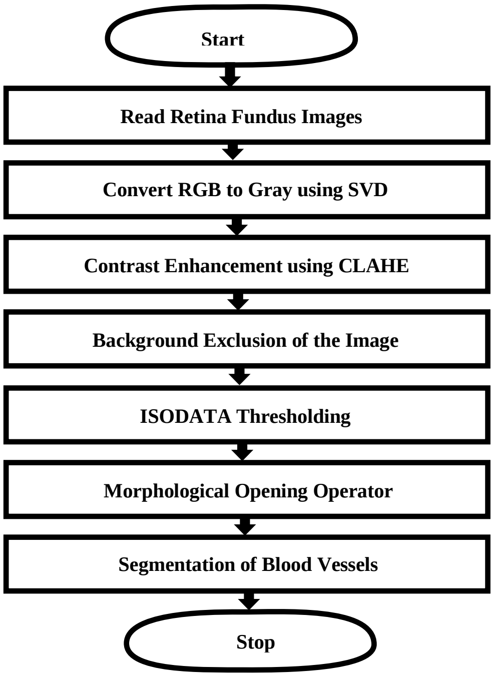

Start
Read Retina Fundus Images
Convert RGB to Gray using SVD
Contrast Enhancement using CLAHE
Background Exclusion of the Image
ISODATA Thresholding
Morphological Opening Operator
Segmentation of Blood Vessels
Stop

Fig. 1. Proposed Method's Flowchart.

a) Firstly, it can be utilized as a processthat transforms interrelated variables or objectsinto a set of unrelated objects or variables that are used to expose a variety of relationships that exists within data under consideration.

b) Secondly, SVD can be used as a perfect means for recognizing and sorting the dimensions of those data elements which show the most deviation.

c) The third form of using SVD is to best approximate the capturedand most deviated data points by means of smaller number dimensions. Thus, SVD could be a best scheme for data reduction.

From the above three primary uses of SVD, this paper uses the second benefit in form of left singular vector matrix to identify and arrange the dimensions(column vectors of the image) that reveal the deviations of blood and non-blood vessels. For reader's perception, a detailed discussion on the background of matrix algebra and vector calculus is followed next.

B. Matrix Algebra and Vector Calculus of SVD

Matrix algebra and vector calculus are two very important mathematical areas for data analysis as well as for statistical theory. Here, this part of the paper deals with introducing the background of vectors and matrices that are applied perfectly to the digital image processing because an image is a 2D function which is a matrix or a two-dimensional array representation in computer memory. Vectors are a series of numbers along various dimensions where those numbers which comprising the vectors are known as components and these can be the measurements of a real-world object in consideration. The components' count represents the dimensionality of that vector. The intuitive in mathematical notation is that the vector V with m-dimensions is a series of m-numbers as such the component v_i is the value of the vector V on i^{th} -dimension. Suppose matrix A is a table of rows and columns which does have two subscripts i.e. n, m which are the maximum count of rows and columns respectively for that matrix. The matrix's entries are called 'components' or 'elements' which are represented by a lower-case letter $'a'$ and the specific entry of the matrix can be located by means of row index which is labeled as i and the column index which is labeled as j . Moreover, an identity matrix can be defined as a square matrix having elements on the diagonal as 1 with remaining other entries as 0.Let the RGB input image under consideration be matrix A possessing various image intensities as column vectors. This image matrix A is transformed into l*a*b* color representation upon which SVD is then computed resulting in three distinct matrices, namely:

a) Left singular vector matrix

b) Diagonal matrix

c) Right singular vector matrix

Mathematically, this is shown in equationbelow

$$A_{mn} = U_{mn} S_{mn} V_{nn}^{T} \tag{1}$$

where, A is the matrix with 'm' rows and 'n' columns' which can be decomposed as the product of three matrices namely (i) matrix U is Left Singular Vector matrix with same order of matrix A (ii) Matrix S is the Diagonal matrix with same order of A (iii) Right Singular Vector matrix is with 'n' rows and 'n' columns.

As this paper made use of the *left singular vector matrix,* the only focus is kept on the discussion of left singular vector matrix.

C. Left Singular Vector Matrix

A matrix U is an orthogonal matrix that is computed on a given input image matrix A and possesses column vectors which areeigenvectorsof AA^T and these vectors are orthonormal.

The detailed explanation of the terms involved in the definition of left singular vector matrix is followed:

D. Orthogonal Matrix

Any matrix is said to be orthogonal matrix if it satisfies the following criteria:

a) Orthogonal matrix should be a square matrix with same count of rows as well as columns. Hence the left singular matrix has rows and columns

b) Matrix U is orthogonal if

$$UU^T = U^T U = I \tag{2}$$

Identity matrix is termed to be a square matrix with component entries on its diagonal are all equal to 1 remaining all other component entries as zero.A matrix A is orthogonal matrixif.

$$A^T = A^T A = I \tag{3}$$

Similarly, a diagonal matrix A contains nonzero values in the main diagonal from its upper left towardsits lower right corner compulsorily and includes all other elements as zeroes. The square matrix's determinant is such a function that condenses that matrix to a particular single number.

E. Orthonormal

Vectors having a length of 1 and are perpendicular with each other are orthonormal. This single word is a combination of two distinct words orthogonal and normal. Any two vectors either insquare or in a rectangular matrix are supposedto be orthogonal (or perpendicular) if their inner product is zero. Such vectors are perpendicular to each other which mean the angle between them is 90^{θ} . Simultaneously, a unit or normal vector is the one whose length is 1. Any vector with a length greater than zero can be made normalized by simply dividing every element by its length. Thecalculation of vector's length is done by squaring each element of the vector, addingall of them, and then computing the square root value of the sum.For example, if we consider $\vec{v}$ as a vector then its length denoted by $|\vec{v}|$ is computed using the equation given below:

$$|\vec{v}| = \sqrt{\sum_{i=1}^{n} X_i^2} \tag{4}$$

The addition of vectors can be computed by adding corresponding components or elements of similar positions in both vectors resulting in a new vector. In general representation, the vector's addition is given by the equation (3) which is stated below:

If $A=[a_{1,}a_{2,}........,a_n]$ and $B=[b_{1,}b_{2,}........,b_n]$ then

$$A+B=[a_1+b_{1,}a_2+b_{2,}........,a_n+b_{n,}] \qquad (5)$$

A scalar value is any real number can be multiplied with a vector implies multiplying that scalar value to every component or element of that vector which produces a new vector. Scalar value multiplication means if d can be any realnumber and $\vec{v}$ can be any vector containing elements as $[x_{1,}x_{2,}........,x_n]$ then a new vector is yielded as shown below in equation (4)

$$d*\vec{v}=[dx_{1,}dx_{2,}........,dx_n] \qquad (6)$$

Two vectors having same dimensions can be multiplied with each other and the result is known as inner product. This can bedone by multiplying corresponding component of one vector by the same positioned component of other vector and adding all of them together is yielding a scalar value. When this scalar value is zero, then the two vectors involved in multiplication are perpendicular or orthogonal to each other. This inner product is also called as scalar or dot product. The two vector's inner product is denoted as

$(\vec{v}_1,\vec{v}_2)$ or $\vec{v}_1,\vec{v}_2$ and generally denoted as below equation:

$$(\vec{X},\vec{Y}) = \vec{X}.\vec{Y} = \sum_{i=1}^{n} X_i Y_i \qquad (7)$$

Eigenvectors

An eigenvector is a vector that is non-zero and fitsin the following equation (6):

$$A\vec{v} = \lambda\vec{v} \qquad (8)$$

where A represents a square matrix**,** λ denotes a scalar value or real number**,** $\vec{v}$ is the eigenvector. By taking matrix as an arrangement of linear equations, the eigenvalues and eigenvector can be found by cracking for those variable values that solely gives the elements of the eigenvector.

Hence, from above discussions upon the theoretical concepts of matrix algebra and vector calculus, it is very elegant and evident that these mathematical techniques play a vital role in image processing, especially, in the preprocessing phase for extraction of the required feature set from given input image.

F. The RGB Color Representation

For proper representation upon computer display, various proportions of red color, green color and blue color are combined to obtain different colors that are in the horizons of visible spectrum. The percentage range of levels of intensities of these three dominant colors is 0 to 100. The color modelRGBcan be manifested with the following intensity function:

$$I_{RGB} = (F_R, F_G, F_B) \qquad (9)$$

where $F_R(x,y)$ corresponds to the pixel's intensity in the red component, $F_G(x,y)$ and $F_B(x,y)$ symbolizes the pixel's intensity in green color as well blue components respectively.The words in [6]mentioned that the color component values get stored in the computer as integer numbers that span 0 to 255 ranges; this is the range that is offered by a single byte which is regularly represented as decimal or hexadecimal numbers. RGB values get encoded as 8-bit integer numbers, which is of 0 to 255 ranges.The intuition here is that the three colors takes 24-bit format meaning that of 24-bit format distributed as 8 bits each for red, for green,for blue. Based on the choice, an adjustment can be made in the quantity of the three portions to derive any of the other colors. For example, if one considers the decimal code for RGB colors, it takes triplet form such that red color in that triplet takes a value of 255 and other two colors as zeroes. This type of decimal coding is applied for green as well for blue color. In order to generate a black color, the decimal code triplet values are all zeroes, for white color the triplet values are all 255 and for gray it is set as (128,128, 128). In this paper, images with high-resolution are taken as input for blood vessels segmentation.

G. The Gray Image

A gray scale image is a matrix with integer numbers as pixel valuesbased on the light intensities of an image that correspond to either black or white color. A grayscale image in MATLABgets stored as an individual matrix where every element in such matrix corresponds to the pixel of an image. The author in [7] discussed that the image's constructionis done using sophisticated sensors and with other image acquirement equipment that signify the brightness (otherwise called as intensity) I of image's light as 2Dfunction which is continuous and is represented as $F(x,y)$ for which (x,y) stand forspatial attributes and whose value corresponds the light's brightness. The matrix can span from class uint8 to double where class uint16, class int16, class single stand between them. The values between [0,255] for uint8, for uint16 [0, 65535],and [-32768, 32767] for int16 are the respective class ranges. For a single or double matrix class,while using the grayscalecolor map, black is represented by 0 intensityand white is represented with intensity 1. For matricesspanning type uint8 to int16,black is represented with intensity intmin(class(I)) and representation of whitewith intensity intmax(class (I)). In this paper, a mask is tailoredbased on Singular Value Decomposition which is used in getting a grayimage from RGB as a preprocessing step.

*H. l*a*b* Color Model*

According to [8], as for identifying a site location having geographic coordinates such as longitude, latitude, and altitude is very similar to have a way to locate and communicate colors using l*a*b* color values where the letter l represents lightness, the letter a represents red/green, and the letter b indicates blue/yellow. During 1940's, Richard Hunter who commenced a tri-stimulus model famously called as *Lab* have given a scaling to attain close unvarying gap of apparentdifferences in color. The *Lab*was made as a benchmark for perfectly plotting complete coordinates of colors and their differences. The l*a*b* or *Lab* color model indicates the color values for the input image. In this work, color space l*a*b* is used as an intermediary model to make a perfect inline in the process of conversion to gray image from input RGB image.

I. Converting into Gray Image from RGB Image

Gray scaled images havecontinuous range of gray values while a binary image has only two possible values for each pixel. Grayscales are represented as integers within the computer. A clear presentation in [9] has given factsabout the pixel value's luminance of animage whose values ranges between 0 to 255. Here, the value 0 denotes black, 255 for white and the values in between 0 and 255 takes the various shades of gray.The process of conversioninto a gray image from RGB image is typically changing the values of RGB (which are of 24-bit) into values of grayscale (which are of 8-bit). The RGB image contains three components and can be taken into consideration as three different images having three scales as red, green, blue that are stacked on top of each other. The array of the orderM*N*3 of color pixel is for RGB image, whereas agrayscaled image is a single layered image that takes shape as M*N array whose pixel values represents intensities. A readily available off-the-shelf function in MATLAB called 'rgb2gray()'iscommonly used to changea RGB image to grayimage. In order to save a single-color pixel of an RGB, we would require 8*3 = 24 bits (8 bit for each color component), but during the conversion to grayscale image from RGB image, only 8-bits are required to preserve an image's single pixel.

III. Related Work

In the literature, a handful of algorithms were proposed to convert into 8-bit gray image from 24-bit RGB image. According to the paper [10], the authors have pioneered to preserve the contrast, sharpness, shadow, and the structure of the image which performs approximation, addition, reduction of luminance and chrominance. The algorithm which was implemented in MATLAB had extracted the three color components into three 2-D matrices and had accumulated the weighted pixel values at corresponding positions of 2-D matrices into a new matrix of the image size. This new matrix holding the weighted sum of each pixel of three color components formed the gray image. The author Tyler Coye in [11] had proposed a new technique which he developed in MATLAB to convert RGB image to gray where he used Principal Component Analysis (PCA) technique to compute principal components or vectors of the input image and then created a weighted mask that got applied on PCA transformed image for convertinginto gray image from RGB image.In the paper [12], the authors proposed a conversion technique depending upon the principle to preserve color image's visual perception. The method proceeds focusing on the (i) *chromaticity contrast*distinctiveness of pixels in the color image rather thanthe luminance contrast, (ii) *Region-based contrast* which instead of individual pixels works upon the regions of the image (iii) Distance-dependent contrast convey the meaning that there should be inverse proportion of the distance between two regions (iv) *Saliency Preservation where s*aliency(features) of the color image needed to properly preserved in the resulting grayscale image. This method had achieved the required aforementioned objectives by formulating anaugmented mapping function for color-to-gray that targets to map each pixel's color of the input image I to the grayscaled value to form the resulting gray image which is given below:

$$gray_p = L_p + C_p \tag{10}$$

where $gray_p$ represents the resulting gray color value of the pixel P in I , L_p is $P's$ value of luminantwhich is independent of the locationis calculated using the below mentioned equation:

$$L_p = L_p^* + (0.1340q(\theta) + 0.0872K_{Br})S_{uv}L_p^* \tag{11}$$

where $q(\theta)$ is the quadrant metric, S_{uv} is the chromatic saturation , and K_{Br} is a constant.

C_p is the $P's$ value of the luminant which is not only location-dependent factorbutalso plays a vital role in adjusting L_p accordingly upon the color variations of P which is calculated using the below mentioned equation:

$$C_p = k\sum_{i=1}^{n}(a_p - a_q)e^{\frac{-\Delta D_{pq}}{\sigma^2}} + l\sum_{i=1}^{n}(b_p - b_q)e^{\frac{-\Delta D_{pq}}{\sigma^2}} \tag{12}$$

where, a_p and a_q are the A channel(the authors of the paper got motivated by chromaticity-based contrast and have measured the differences in chrominance by a combination of two distance-weighted color differences of two channels where they named them as A channel and B channel of CIELAB color space) values of pixels p and q respectively. Similarly, b_p and b_q are the values of B channel corresponding to and q . D_{pq} is the Euclidean distance between p and q . The exponent e is used to approximate the location-dependent contrast. Also, σ^2 is been set with an fuzzy value between 0 and 1 depending on the width and height of the image. Here, k and l are variables which are unknown and have been optimized depending on the criteria of saliency preservation. n is the size of the pixels in the image.

In this paper, the approach followed is very specific where a mask is created upon the calculated left singular value matrix of the l*a*b* transformed input RGB image. The mask used here operates upon every organized pixel of Left Singular Vector matrix to convert them to take values of gray. This new technique has been developed using MATLAB software. The steps involved in preprocessing the image which is performed after the input image is read, resized for easy computation, and get doubled which holds real values between 0 to 255 of the image, are mentioned below:

Step 1 Translate the doubled image to l*a*b* color model.

Step 2 Compute the new image with the weighted combination of the vectors of the resultant l*a*b* model image of Step (1) in order to increase the contrast in the gray version of those partswith different hues but having similar intensities.

Step 3 Decompose the image of Step (2) using SVD into three matrices namely: Left Singular Vector Matrix, Diagonal Matrix, and Right Singular Vector Matrix.

Step 4 Extract Left Singular Vector Matrix of the SVD which holds the feature set in column vectors of decreasing order.

Step 5 To have the size compatibility, reshape Left Singular Vector Matrix to fit to the size of resultant image of Step (1) such that the required feature set can be accessed as the first portion with that size.

Step 6 Access the featured column vectors and normalize it with maskthat works on the every pixel of the column vector. The mask operates on each pixel where at each pixel the computation takes place as the ratio of the computed result of subtraction ofpixel of the matrix with the least pixel with thecomputed result of subtraction of maximum pixel with least pixel of the matrix. The outcome of this step is the gray version of the given input RGB image.

The gray scaled image, the outcome of the conversion procedure disclosed above, is fed as the input for contrast enhancement to the CLAHE which is discussed next.

A. AbbreviationsAdaptive Histogram Equalization (AHE)

A perfect analysis of an image needs a perfect pre-processing of it using contrast enhancement technique which can be done locally or globally. The study done by [13] revealed that the usage of CLAHE which is based on local contrast enhancement had given optimal results which when compared to other global enhancement schemes namely histogram specification, BSB-CLAHE, histogram equalization, AHE that mostly get applied on colored image segmentation. The authors in paper [14] have utilized the same CLAHE along with other Wiener filter method to enhance the image's contrast. Usually, they used filters to take out the noise there in image and to shift image into appropriate dynamic range they used Gamma correction techniques. Finally, they went with CLAHE that avoided amplification of unnecessary noise hidden in the image and have devised its parameters to constrain the contrast in identical areas. Image sharpness can be increased by magnifying image's contrast which is one of the important tasks in a system of automated disease diagnosis.Contrast Limited Adaptive Histogram Equalization (CLAHE) is animage processingtechnique which is an extended edition of adaptive histogram equalization (AHE). The computerized procedure AHE is familiar in the realm of image processing which is primarily used to amplify the image's contrast. This technique which works on small subdivisions rather than whole image, initially, computes histogram for every region and then redistributes the lightness values to that particular region. Despite making improvements in the image's local contrast, this method has huge chances to make strong the noise in those regions that are relatively homogeneous of that image. To overcome this, CLAHE is used to act globally and limiting the contrast. In this paper, pre-processing of the image is done using CLAHE which takes SVD transformed gray image as a parameter and ideally enhances the contrast of those relevant features which are being captured by left singular SVD matrix.

B. Averaging Filter for Smoothing

To reduce the intensity variation among pixels and to reduce noise that is present in the input image, filters are extensively used and are the facilitators in the domain of processing of the image.Mean filter or an average filter is windowed and linear class filters that is easy to implementto smooth the given image. Such filters usually works as best low-pass one. For any element of animage, take an average across its neighborhood based on the size of the mask chosen, for say, 3x3 or 5x5 or 9x9, and replace the average value to that location of the element is a very simple idea behind this mean filter.

C. ISODATAThresholding for Segmentation

The simplest segmentation method in image processing is thresholding. This method splits the given image into minor segments using a single color scale named gray scale value for properly defining boundary of each. The advantage behind getting a binary image is realized in decreasing the given input data'sdifficultyand extensively simplifying the recognition of the object and its classification processes [15]. The pixels get partitioned based on their intensity value.Once having thegrayscale image as a parameter to this method, then thresholding easily creates images that are binary. The process of this technique moves by considering each individual pixelthat isscored based on the value compared to the desired threshold value. The pixel is an "object" pixel if the score is greater than the given threshold and is "background" pixel if the score value falls below the threshold value. During this process of deciding on whether a pixel is object or background, the replacement is made with 1 if it is an object pixel and with 0 if it is a background pixel. ISODATA classification scheme computes the class 'means'of the data space that is evenly distributed in it. Then, it proceeds with an iterative approach to cluster all other remaining pixels with the help of techniques that calculates minimum distance. This algorithm recalculates 'means'at every iteration and reclassifies all other remaining pixels accordingly aligned with the new means being computed. Based on the input threshold parameters, this technique performs iterative splitting of the class, merging, and deletions if required. Here, classification of every pixel to its nearest class is done unless specification is made for a standard deviation measure or for the value of the distance threshold. If went unspecified, some pixels may have the chances to get forego classification once the selected criteria is not met by them. This procedure iterates until pixels countin every classgets changedwhich should be at most the value of the chosen pixel change threshold or until the maximumiterations specified is reached [16].

In this paper, ISODATA(Iterative Self-Organizing Data Analysis Technique) thresholding is performed by taking resultant subtracted image that is being computed between gray image and mean filter image as an input parameter.

D. Morphological Area Open Operation

According to [16], theoperations that span a broad set in the domain of image processing as well as works upon the object'sshapes in the given input image is carried out by Morphological Operators. These operations of the operators which apply a structuring element to createa desired and same sizedoutput image are very much familiar to the researchers who witness their effort in segmentation of images. The operation of these operators depends on getting the corresponding pixel value in the output image where each output pixel results from the input pixel value after being subject to work on with its neighboring pixel values based on the given criteria. Mostly on binary images these morphological operators are used, e.g., for background subtraction from an image. They are used on gray value images, if they are viewed as a stack to binary images**.** Most of the operations used in image processing are a combination of two processes, dilation and erosion. Moreover, the two forms of structuring element which work on binary images and gray images respectively are flat and non-flat. In the paper [17], morphological area open is used toremove all of the connected components that do have fewer pixels than the given parametric pixel size from the binary image which is the offset parameter to this method which outputs in a binary image as the final segmented image.The morphological opening is very useful for conserving the object's shape and size along while removing small unwanted objects. The mathematical formula for the morphological opening is given below in equation (11) which is the calculation of dilation from the calculated erosion where erosion works upon a set of 'A' by using a structuring element 'B'which is defined as

$$AoB = (A\square\;\; B) \oplus B \tag{13}$$

where $\square$ denotes erosion and $\oplus$ denotes dilation.

In this proposed work, the morphological opening operator has been used to remove small-sized and unwanted blood vessels while safeguarding the blood vessels shape and size. Here, the function takes two parameters: (i) Flat structuring element parameter: the value of this parameter is the binary image which is computed after the ISO thresholding method discussed above (ii) Offset length: the value of this parameter is an integer number which is set between 20 and 100 for the images being worked on. This offset length is crucial as the removal of the blood vessels depends upon the value been set.

IV. Results and Discussions

A. STARE and DRIVEDatabases

*1) STARE database:*There is another database containing retina color fundus images by name STructured Analysis of the REtina (STARE)[19].The training dataset relating to blood vessels segmentation contains 40 manually segmented images performed by two observers and are subdivided into two sets where the first set is termed as first observer and the second one as second observer. Test dataset have 20 color images: 10 usual and 10unusual which are beingtaken by a TopCon TRV-50 fundus camera atFOV of 35^{o} and the image's size is 700 X 605 pixels.

*2) DRIVE database:*DigitalRetinal Images for Vessel Extraction (DRIVE) database [18] is such a database introduced in the year 2004 to serve the perfect purpose in carrying out an ideal research work in the field of retinal fundus image processing.There are about 40 retina images among them 20are training images and other 20 are test images which are being captured by non-mydriatic3CCDCanon CR5 camera at 45_ field of view (FOV). Theimage's size is of 768 x 584 pixels where it takes 8-bits to each colorchannel. Also, this database contains both normalas well as abnormal images along with manuallylabeled images by experts to make an assessment of the proposed method‘s performance.The most important aspect of this database is the inclusion of the masks for separation of FOV from image's remaining part.Another aspect of this database is regarding theavailability of two hand labeled setsin which thefirst set provides the ground truth for results evaluation of any proposed manual while the second set containshalf of the images which are hand labeled.

B. Performance Measures

The method which is proposed in this paper has been implemented inMATLAB 2017Bb andthe assessment of the performanceisdone in view of its specificity, sensitivity, and accuracy. The proposed method is implemented on DRIVE and STARE database images. This method has utilized the Singular Value Decomposition (SVD) for effective conversion of RGB to gray, followed by adaptive histogram equalization (AHE) for good enhancement results. After having a good contrast enhancement, the effective segmentationis followed by ISODATAthresholding. Here, the thresholded segmentation results havesuffered from some boundary leakages and unwanted pixels presence. In order to remove these drawbacks of thresholding method, the morphological opening is used to prune away the small and needless blood vessels finally producing a best segmented image as output.The metrics used in analyzing the performance of the proposed technique are:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{14}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{15}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

True Positive (TP) give the count of pixelsin the image being segmented and also in the groundtruth.False Positive

(FP) gives the count of pixels detected in the image under consideration but not inground truth of the same image.Trueratio is defined as the ratio of the count of true pixels belongs to blood vesselsand the countof ground truth's vessel pixels. Finally, falseratio is the ratio of non-blood vessel pixels count to thecount ofground truth's non-vessel pixels.

C. Performance Comparison of Existing and Proposed Methods

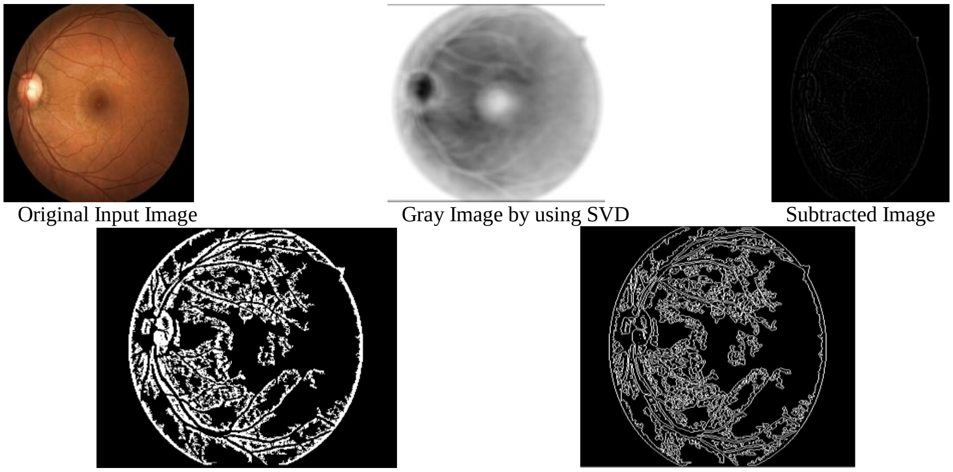
The segmentation performance of the proposed method is analyzed based on the parameters such as sensitivity, specificity and accuracy. The performance measures are calculated based on the equations (12, 13 and 14). The total count of the pixels properly classified as the vessel (true positives) pixel and non-vessel pixel(true negatives) divided by the total count of pixels in the images determines the accuracy of the segmentation algorithm. Moreover, the measurement of the segmentation correctness is obtained through true positive rate (TPR) and false positive rate (FPR). TPR is the result of the ratio of total count of true positives to the count of true blood vessel pixels. The FPR is the outcome of the ratio of the total count of false positives to the count of non-vessel pixels spotted in the ground truth image.

The performance measures of the existing blood vessels in retinal images as shown in the Table I. From the Table I, first eight rows show the performance of existing segmentation for blood vessels. The maximum segmentation accuracy of existing methods is such as Youetal (2011),Kumar et al (2016), Soares et al (2006),Mendonca et al (2006) , B. Zhang et al (2010), Martinez-Perez et al (2007), Mendonca and Campilho et al (2006), Palomera-Perez et al. (2010)were achieved an accuracy of 94.65%,96.26%,94.80%,94.66%, 93.82%, 93.44% , 94.52% and 92.55% respectively.

In the literature survey on blood vessels segmentation algorithms, it is found that the pre-processing stage plays a crucial role to get the effective segmentation accuracy. So, the proposed algorithm does the pre-processing of images by using singular valued decomposition (SVD) for RGB to gray scale effectively. This algorithm is tested on four different images that are taken from databases mentioned. After successful testing and assessment, the outstanding performance is witnessed in the parameters such as sensitivity, specificity and accuracy of the proposed method tabulated in last row of Table I. Moreover, segmentation accuracy of the proposed method is 97.98%, 98.45%, 98.32% and 98.78% of four images as shown in Fig. 2, 3, 4 and 5, respectively. Finally, based on the performance analysis done in the Table I, the proposed method segments the retinal blood vessels accurately and effectively of 97.48% compared with the existing blood vessels segmentation algorithms.

Fig. 2, 3, 4 and 5 also show the results of proposed method, Fig. 2(a), 3(a), 4(a) and 5(a) are the input retinal blood vessels fundus images, Fig. 2(b), 3(b), 4(b) and 5(b) are the effective RGB to Gray conversion images by using SVD, Fig. 2(c), 3(c), 4(c), and 5(c) are the subtracted images of gray and average filter, Fig. 2(d,e), 3(d,e), 4(d,e) and 5(d,e) are thefinal blood vessels segmentation and its perimeter images respectively. The result analysis of the proposed approachis provided in Table I. Table II gives the segmentation area of the blood vessels covered by the proposed method.

D. Experimental Results on Various Blood Vessels in Fundus Images



Original Input Image

Gray Image by using SVD

Subtracted Image

Final Segmented blood vessels using Proposed method

Perimeters of the segmented blood vessels

Fig. 2. Segmentation Results of the Proposed Method based on SVD and Morphological Opening on Retinal Images: (a) Depicts the Input Retinal Images with Blood Vessels, (b) Shows the Contrast Enhancement Image using SVD, (c) Depicts the Subtracted Image of Gray an Average Filter, (d) and (e) Shows the Final Blood Vessels Segmentation and its Perimeter of the Resultant Image Respectively.

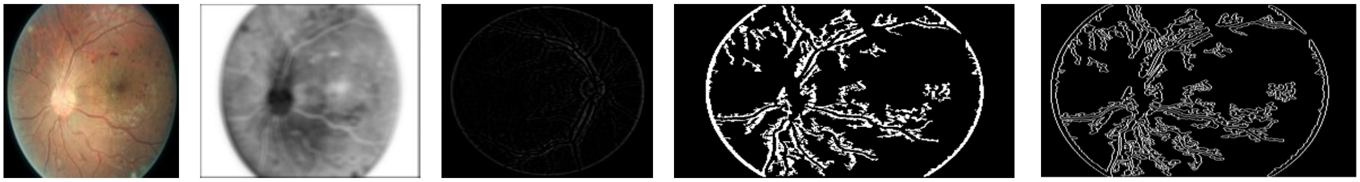


Fig. 3. Segmentation Results of the Proposed Method based on SVD and Morphological Opening on Retinal Images:(a) Depicts the Input Retinal Images with Blood Vessels, (b) Shows the Contrast Enhancement Image using SVD, (c) Depicts the Subtracted Image of Gray and Average Filter, (d) and (e) Shows the Final Blood Vessels Segmentation and its Perimeter of the Resultant Image Respectively.

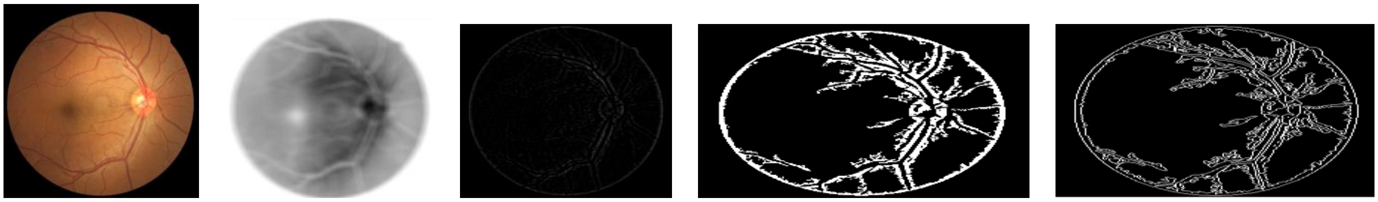


Fig. 4. Segmentation Results of the Proposed Method based on SVD and Morphological Opening on Retinal Images: (a) Depicts the Input Retinal Images with Blood Vessels, (b) Shows the Contrast Enhancement Image using SVD, (c) Depicts the Subtracted Image of Gray and Average Filter, (d) and (e) Shows the Final Blood Vessels Segmentation and its Perimeter of the Resultant Image Respectively.

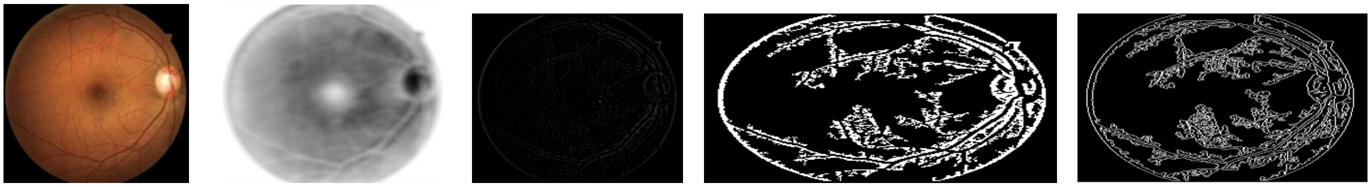


Fig. 5. Segmentation Results of the Proposed Method based on SVD and Morphological Opening on Retinal Images: (a) Depicts the Input Retinal Images with Blood Vessels, (b) Shows the Contrast Enhancement Image using SVD, (c) Depicts the Subtracted Image of Gray an Average Filter, (d) and (e) Shows the Final Blood Vessels Segmentation and its Perimeter of the Resultant Image Respectively.

TABLE. I. PERFORMANCE COMPARISON BETWEEN EXISTING METHODS AND PROPOSED METHOD IN TERMS OF SENSITIVITY, SPECIFICITY, AND ACCURACY

No	Author/Authors	Database	Sensitivity	Specificity	TPR	FPR	Accuracy	Technique used
1	You et al (2011) [20]	DRIVE	0.7410	0.9751	-	-	0.9434	Radial projection and semi-supervised approach.
		STARE	0.7260	0.9756	-	-	0.9497	
2	Kumar et al (2016) [21]	DRIVE	0.7006	0.9871	-	-	0.9626	Laplacian 2-D Matched Filter
		STARE	0.7060	0.9693	-	-	0.9340	
3	Soares et al (2006) [22]	DRIVE	-	-	0.7165	0.0252	0.9480	MF-FDOG
		STARE	-	-	0.7283	0.0212	0.9440	
4	Mendonca et al (2006) [23]	DRIVE	-	-	0.6996	0.0270	0.9466	MF-FDOG
		STARE	-	-	0.7344	0.0236	0.9452	
5	B. Zhang et al (2010) [24]	DRIVE	0.7120	0.9724	-	-	0.9382	Matched Filter with first order derivative of Gaussian
		STARE	0.7177	0.9753	-	-	0.9484	
6	Martinez-Perez et al (2007) [25]	DRIVE	0.7246	0.9655	-	-	0.9344	Multi-scale feature extraction
		STARE	0.7506	0.9569	-	-	0.9410	
7	Mendonca et al (2006) [26]	DRIVE	0.7344	0.9764	-	-	0.9452	Center lines detection and morphological reconstruction
		STARE	0.6996	0.9730	-	-	0.9440	
8	Palomera-Perez et al. (2010) [27]	DRIVE	0.64	0.967	-	-	0.9250	Parallel multiscale feature extraction and region growing
		STARE	0.769	0.9449	-	-	0.926	
9.	**Proposed method**	STARE (Fig.2)	**0.9123**	**0.9556**	**0.8978**	**0.0092**	**0.9738**	**Singular Valued Decomposition (SVD)**
		STARE (Fig.3)	**0.9443**	**0.9567**	**0.9367**	**0.0008**	**0.9745**	**SVD**
		DRIVE (Fig.4)	**0.9468**	**0.9666**	**0.9489**	**0.0007**	**0.9732**	**SVD**
		DRIVE (Fig.5)	**0.9567**	**0.9687**	**0.9567**	**0.0005**	**0.9778**	**SVD**

TABLE. II. BLOOD VESSELS SEGMENTATION AREA COVERED BY USING PROPOSED METHOD

Proposed Method/ Blood vessels Images	**Image 1 (Fig. 2)**	**Image 2 (Fig. 3)**	**Image 3 (Fig. 3)**	**Image 4 (Fig. 4)**
Blood vessels segmentation area covered	474.4525 mm^2	504.8308 mm^2	519.9923 mm^2	496.3786 mm^2

V. CONCLUSION AND FUTURE WORK

In this paper, a novel algorithm to segment tree-like vascular structure from retinal images is proposed. The procedure has been successful to pull out the required object features of an input image using the SVD's left singular vector matrix for transformation into gray image. This method has been practically proved to be effective in attaining average segmentation accuracy of 97.48%, which is superior compared with existing segmentation algorithms listed in Table I. This new technique got tested upon images contained in DRIVE and STARE databases. In order to attain further accuracy, this technique can be combined with any of the optimization techniques. This combination of the proposed method and an optimization technique will give best results.

ACKNOWLEDGMENT

I thank my guide, faculty, Head of the Department, Principal of GITAM Institute of Technology (GIT) for their kind support and encouragement to carry out this work.

REFERENCES

[1] G. J.J. Kanski, Clinical Ophthalmology, 6th ed., ElsevierHealth Sciences, London, UK, 2007.

[2] K. Verma, P. Deep, A.G. Ramakrishnan, “Detection and classification of diabetic retinopathy using retinal images flight’, Annual IEEE India Conference (INDICON), vol. 4, pp. 1–6, 2011.

[3] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A.R. Rudnicka, C.G. Owen, S.A. Barman, “Blood vessel segmentation methodologies in retinal images a survey”, Comput. Methods Programs Biomed, vol. 1, pp. 407–433, 2012.

[4] T. Walter, P. Massin, A. Erginay, R. Ordonez, C. Jeulin, J.C. Klein, Automatic detection of microaneurysms in color fundus images, Med. Image Anal. 11 (2007) 555–566.

[5] D. Kahaner, C. Moler and S. Nash, Numerical Methods and Software, Prentice-Hall, Inc, 1989.

[6] https://en.wikipedia.org/wiki/RGB_color_model.

[7] Tarun Kumar, Karun Verma “A Theory Based on Conversion of RGB image to Gray image” International Journal of Computer Applications (0975– 8887) Volume 7– No.2, September 2010.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[8] https://www.xrite.com/blog/lab-color-space.

[9] Saravanan, C. (2010). Color Image to Grayscale Image Conversion. 2010 Second International Conference on Computer Engineering and Applications.doi:10.1109/iccea.2010.192.

[10] T. Walter, P. Massin, A. Erginay, R. Ordonez, C. Jeulin, J.C. Klein, Automatic detection of microaneurysms in color fundus images, Med. Image Anal. 11 (2007) 555–566.

[11] C. Tyler, “A Novel Retinal Blood Vessel Segmentation Algorithm for Fundus Images,” MATLAB Cent. File Exch., 2016.

[12] Du, H., He, S., Sheng, B., Ma, L., & Lau, R. W. H. (2015). Saliency-Guided Color-to-Gray Conversion Using Region-Based Optimization. IEEE Transactions on Image Processing, 24(1), 434–443.doi:10.1109/tip.2014.2380172.

[13] DibyaJyoti Bora, “Importance of Image Enhancement Techniques in Color Image Segmentation: A comprehensive and Comparative Study”, Indian J.Sci.Res. 15, vol. 4, pp. 115-131, 2017.

[14] Mithilesh Kumar, AshimaRana, “Image Enhancement using Contrast Limited Adaptive Histogram Equalization and Wiener filter”, International Journal Of Engineering And Computer Science, vol. 5, pp. 16977-16979, 2016.

[15] K. Bhargavi, S. Jyothi, “A Survey on Threshold basedSegmentation Technique in Image Processing”, International Journal Of Innovative Research and Development, (Vol 3 Issue 12, November, 2014.

[16] https://www.harrisgeospatial.com/docs/ISODATAClassification.html.

[17] https://in.mathworks.com/help/images/morphological-dilation-and-erosion.html.

[18] Niemeijer M, Staal JJ, GinnekenBv, Loog M, Abramoff MD.DRIVE: digital retinal images for vessel extraction; 2004. WebLink:http://www.isi.uu.nl/Research/Databases/DRIVE/.

[19] The STARE project, [online] Available: http: //www.ces.clemson.edu/~ahoover/stare.

[20] You, X., Peng, Q., Yuan, Y., Cheung, Y., & Lei, J. (2011). Segmentation of retinal blood vessels using the radial projection and semi-supervised approach. Pattern Recognition, 44(10-11), 2314–2324.doi:10.1016/j.patcog.2011.01.007.

[21] Kumar, D., Pramanik, A., Kar, S. S., &Maity, S. P. (2016). Retinal blood vessel segmentation using matched filter and Laplacian of Gaussian. 2016 International Conference on Signal Processing and Communications (SPCOM). doi:10.1109/spcom.2016.7746666.

[22] J.V.B. Soares, J.J.G. Leandro, R.M. Cesar-Jr., H.F. Jelinek, and M.J. Cree, “Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification,” IEEE Trans. on Medical Imaging, vol. 25, pp. 1214–1222, 2006.

[23] Mendonca, A. M., &Campilho, A. (2006). Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. IEEE Transactions on Medical Imaging, 25(9), 1200–1213. doi:10.1109/tmi.2006.879955.

[24] B. Zhang, L. Zhang, L. Zhang, F. Karray, Retinal vessel extraction by matched filter with first-order derivative of Gaussian, Computers in Biology and Medicine 40 (2010)438–445.

[25] M.E. Martinez-Perez, A.D. Hughes, S.A. Thom, A.A. Bharath, K.H. Parker, Segmentation of blood vessels from red-free and fluorescein retinal images, Medical Image Analysis 11 (2007) 47–61.

[26] Safsfd A.M. Mendonca, A. Campilho, Segmentation of retinalblood vessels by combining the detection of centerlines and morphological reconstruction, IEEE Transactions on Medical Imaging 25 (2006) 1200–1213.

[27] Palomera-Perez, M. A., Martinez-Perez, M. E., Benitez-Perez, H., & Ortega-Arjona, J. L. (2010). Parallel Multiscale Feature Extraction and Region Growing: Application in Retinal Blood Vessel Detection. IEEE Transactions on Information Technology in Biomedicine, 14(2), 500–506.doi:10.1109/titb.2009.2036604.

A Survey of Various Frameworks and Solutions in all Branches of Digital Forensics with a Focus on Cloud Forensics

Mohammed Khanafseh[1], Mohammad Qatawneh[2]
King Abdulla II School for Information and Technology
The University of Jordan Amman-Jordan

Wesam Almobaideen[3]
The University of Jordan Department of Computer Science
Rochester Institute of Technology Dubai, UAE

***Abstract*—Digital forensics is a class of forensic science interested with the use of digital information produced, stored and transmitted by various digital devices as source of evidence in investigations and legal proceedings. Digital forensics can be split up to several classes such as computer forensics, network forensics, mobile forensics, cloud computing forensics, and IoT forensics. In recent years, cloud computing has emerged as a popular computing model in various areas of human life. However, cloud computing systems lack support for computer forensic investigations. The main goal of digital forensics is to prove the presence of a particular document in a given digital device. This paper presents a comprehensive survey of various frameworks and solutions in all classes of digital forensics with a focus on cloud forensics. We start by discussing different forensics classes, their frameworks, limitations and solutions. Then we focus on the methodological aspect and existing challenges of cloud forensics. Moreover, the detailed comparison discusses drawbacks, differences and similarities of several suggested cloud computing frameworks providing future research directions.**

***Keywords*—Digital forensics; cloud forensics; investigation process; IoT forensics; examination stage; evidence**

I. Introduction

Digital forensics refers to the science, which deals with the crimes that happened on the level of digital devices. The main purpose of digital forensics is to detect, extract, and analyze evidence from digital media and prepare it for the prosecution, so that a case can be presented in court [1][2]. Using digital devices as a criminal tool has enhanced the criminal's ability to do different activities of criminals such as hiding facts, updating facts and users' document, or any other unethical activity. This type of crimes, i.e. "cybercrimes" is an extension of classical crimes. To deal with different crimes that happened on the level of digital devices, the investigators must implement consistent and precisely defined forensic procedures.

The Investigation process of any digital crime depends mainly on identifying and collecting evidence from the resources. The digital evidence refers to any critical information relevant to proof of the crime. This information, which can be used and accepted in court, stored and transmitted in digital form [3][4].

Moreover, the investigation process is highly depending on the device type and the environment used, which means that there is more than one branch under digital forensics because digital devices can be traditional computers, mobiles, network devices such routers, etc. Several challenges have been faced by digital forensics which hampers in finding out digital evidences such as technical, legal, and resource challenges. Consequently, the investigation process for one digital device may not be used for the other device so, it is difficult to find out a process which is compatible with all devices and environments.

This paper gives a comprehensive survey of different digital forensics branches which help the investigator to proceed in its investigation process. Moreover, the survey focuses on challenges, frameworks, and solutions of cloud forensics. The rest of this paper is organized as follows. Section 2 presents an introduction to digital forensics; Section 3 discusses the main branches of digital forensics, their frameworks, solutions and drawbacks. Finally, the article is concluded with the outcomes based on many comparisons between different branches of digital forensics and future work.

II. Digital Forensics

The digital forensic process is still in its infancy, but it is becoming increasingly invaluable for researchers, and many researchers have recently been working to propose specific models for digital forensics. The first proposed models for digital forensics include four main stages: Acquisition stage, Identification stage, Evaluation stage, and Admission stage. Since then, many models have been proposed to explain the steps taken to acquire, identify and analyze the evidence obtained from different digital devices. Digital forensics has become commonplace due to the increasing spread of technology and the high level of technological dependency since the 20Th century. In tradition forensics, the evidence is something tangible that could identify the criminal, such as blood, fingerprint, and hair, but these evidences cannot be found at digital forensics. In general, digital word refers to something related to computer technology such as files and data in digital form, based on that digital evidence refer to anything that can be extracted from digital devices. Based on the increasing number of digital devices, and the highly dependent on these devices on daily activity for different persons, the number of digital crimes was increased.

The number of digital devices that require analysis is also increased, and the storage volume for each device is also increased. These devices and their storage space increase the complexity of digital forensics process.

A standard framework which proposed to guide the process of digital forensics is important to accelerate the process of investigation, and to overcome different problems that faces the process of investigation such as huge volume of storage space on digital devices [5],[6]. Several frameworks were defined over the time; each of new frameworks tries to integrate new technology and methods over the previous one. Most of the research in recent years was more concerned with employing new methods and technology to improve current frameworks from different aspects such as from efficiency and accuracy aspects; other current research was concerned to dealing with new problems.

A. Various Definitions of Digital Forensics

Various definitions of digital forensics have been proposed by many researchers, depending on legal, criminal or a process perspective. This section discusses some of these definitions. Some of the proposed definitions are as follows:

1) The researchers in [5] established one of the first definitions of DF. They define a Digital Forensics (DF) as a branch of forensics science involved with the use of scientific techniques towards the preservation, collection, validation, identification, analysis, interpretation, and presentation of digital evidence derived from digital sources for the purpose of facilitating the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned actions.

2) Another definition of digital forensics is suggested by [7], the author defines DF as a widely used term referring to identification, acquisition and preservation, digital evidence analysis from various digital devices, not just computers such as camera, smartphone, tablet, IOT device, network device, and other digital devices.

3) Other researchers, such as in [8] define a simple definition of digital forensics as science for identifying, preserving, recovering, analyzing and presenting facts and evidence relating to digital evidence on digital device storage media. This definition is a good and general definition because it is a simple definition that divides the process of forensics into four main areas from identification to the presentation.

4) Another simple definition has been proposed in [1], which defines a digital forensics as science for detecting, extracting and analyzing evidence from digital media, and is one of the critical requirements in cyberspace. And the author through his definition defines the main purpose of digital forensics through the definition is to prepare a report accepted in a court.

B. Digital Forensics Frameworks

Many researchers have developed a new process models and solutions to improve digital forensics. Digital forensics as science has made significant progress not only in the field of technology but also in methodology improvement. The process model is the methodology used through digital forensics to conduct a research framework with a number of phases that guide the investigation process. Generally speaking, there is no standard framework for the investigation process because the investigation process depends on the area of investigation and on a variety of cases, e.g., cyber attachment by IT specialists, civil cases in a corporation, or criminal cases, so different investigators will follow different ways of investigating their investigation process. The standard method used through conventional digital forensic processes, such as in [9] [10] [11], consists of defining the sequence of dependent stages necessary for the investigation process. The frameworks used during the investigation process can be classified based on a number of stages and sub-stages used. If the framework used contains a few stages, then this framework will not provide much guidance for the investigation process. A framework that contains many stages, and each stage has sub-stages, with its usage scenario being more limited, may prove more useful. Therefore, none of the proposed frameworks can have a general purpose and be used on any type of investigation, but the idea of any proposed framework should be as general as possible, which could be applied to as many cases as possible. To date, various frameworks have been proposed in the field of digital forensics. For a specific case, each of the proposed frameworks attempts to refine a particular methodology, for a particular case, different model have the same steps for the same case. Earliest research on digital forensics focused on defining the process of digital forensics [6].

Recently, the frameworks that proposed in digital forensic focused on a specific stage of digital forensics stages such as (identification, collection, preservation, and examination analysis stages), such as triage framework [12][13], which has been developed for time-sensitive applications. By applying digital forensics triage, the investigator could find the related evidence, which can accelerate the investigation process to lead to the criminal rather than waiting for the whole report from the police which could take several months or even years. Many frameworks were proposed on different areas of digital field, such as in digital forensics in general, computer forensics, mobile forensics, network forensics, IoT forensics, and cloud forensics. Each proposed framework has its own characteristics such as number of stages and strategy used for evidence collection based on the area of implementation.

Digital forensics framework can be defined as a structure to support a successful forensic investigation. This implies that the conclusion reached by one computer forensic expert should be the same as that of any other person who has conducted the same investigation [14]. Standardized digital forensics framework consists from each of the following stages [5].

1) Identification stage: This stage recognizes an incident from indicators and determines its type.

2) Preparation stage, which entails the preparation of tools, techniques, search warrants and monitoring authorizations and management support.

3) Approach strategy that develops a procedure to use in order to maximize the collection of untainted evidence while minimizing the impact to the victim.

4) Preservation stage: The preparation stage involves the isolation, securing and preservation of the state of physical and digital evidence.

5) Collection stage that entails the recording of the physical scene and duplicate digital evidence using standardized and accepted procedures.

6) Examination stage which involves an in-depth systematic search of evidence relating to the suspected crime.

7) Analysis stage which involves determination of the significance, reconstructing fragments of data and drawing conclusions based on evidence found.

8) Presentation stage which involves the summary and explanation of conclusions.

9) Returning evidence that ensures physical and digital property is returned to the proper owner.

Many frameworks have been proposed for investigation digital crimes, where each framework consists of a set of upper stages, and some others contain all standard stages of the digital forensics model. Table I shows different digital forensics frameworks proposed by many researchers.

Table I presents the most common stages of different forensics frameworks, problems solved by each framework and defines the degree of complexity of these frameworks based on the number of branches used. The conclusion that can be reached from Table I is that there are a set of issues that are not taken into account in the proposed frameworks such as confidentiality, security awareness, and accuracy of the investigation process in specific through evidence collection and examination steps. Based on the standard stages of any digital investigation process, the digital forensics frameworks can be categorized into the following three main types:

1) General Frameworks: Such frameworks proposed between 2000 and 2010 and focused on defining the phases of typical investigations [17]. Examples of this type of frameworks are shown in Table II [18][19][20][21].

Table II contains the frameworks which can be classified under the first type of digital forensics. The table contains information about each framework, such as the name of the framework, the main stages of each framework, and more, the table contains the sub-stage for each stage of the framework such as in the identification stage, some frameworks in this type of digital forensics frameworks contain just the stage of identification but in other frameworks, this stage was divided into more than sub-stage such as in Lee framework which divides it into classify sub-stage and compare sub-stage. Furthermore, this table provides a comment row which contains the main comments on the individual frameworks. The strengths and weaknesses of the framework and the complexity of the frameworks depend on the number of main stages and sub-stages in the frameworks proposed.

1) Frameworks concerned with a specific case, and focusing on a particular stage of digital forensics: Different frameworks of this type have been suggested such as frameworks which deal with certain categories of cases like forensic network, computer forensics, IOT forensics and other forensics, and frameworks that proposed for sensitive cases like abductions, missing person cases, etc., [22], [12]. At the early stages of digital forensics science, the proposed digital frameworks faced different issues, one of the urgent issues is to define the process model to make the entire investigation process consistent and standardized, general process model have been defined for investigation process, later framework that proposed contain additions stages for process model and with sub stages for main stages that forms the early framework, many of the new frameworks that have been proposed for digital forensic investigation which depends mainly on early suggested frameworks, Table III shows both original frameworks and updated frameworks, each of new proposed frameworks depends on the stages and strategy of investigation on conventional framework such as the SRDFIM framework which proposed in 2011 depends on the framework which called DFRWS framework, and many other frameworks as mentioned in Table III.

2) Frameworks have been proposed in recent years to deal with new technology such as cloud computing forensics, IOT forensics, etc. Some of the latest technology leads to new problems hampering digital forensic investigations, such as the problems that arise in cloud computing forensics. Cloud computing forensics faces many problems through each stage of the investigation process, such as the difficulty of identifying the right resources through the identification stage, matching the right evidence through examination and collection phases, etc. Other problems are caused by the use of digital forensics in the crime that occurred in the smart environment containing IoT devices, because more digital devices connected to the internet result in an ever- increasing volume of data. Based on these problems faced by forensics on new technology, a new integration between forensics processes and new technologies such as mining algorithms, security algorithms, data integrity and authentication algorithms that proposed to overcome all these problems, new frameworks were proposed in other more sophisticated cases to address the problem that can be solved through integration.

Table IV shows the main frameworks and solutions that were proposed for the IoT environment. The table contains the main stages of each framework, the main idea and goal of each framework, the name of the technology that the framework developed to serve, the encouragement point to develop the framework, the main idea for the proposed framework and what is the enhanced point in the framework and the name of the original framework on which the proposed framework depends if it is found.

TABLE. I. MAIN FRAMEWORKS PROPOSED FOR DIGITAL FORENSICS

Framework Names	Digital Forensic Branch Name	—	Individualization Stage	Initialization Stage	Identification Stage	Collection Stage	Authentication Stage	Preservation stage	Evidence Reduction stage	Documentation stage	Analysis Stage	Examination Stage	Presentation Stage	Reporting	Decision Stage	Review Stage	Plan Stage	Transportation Stage	Complexity Stage	Other Stages	Contribution And Comments
Generic Process Model for Network Forensics	Network Forensics	χ	χ	✓	✓	✓	χ	✓	✓	χ	✓	✓	χ	✓	χ	χ	χ	χ	H	Y	The proposed framework contains main stages of digital forensics with incorporating a new stage for detecting [15]
A Framework for a Digital Forensics Investigation	Digital Forensics	✓	χ	χ	✓	✓	χ	✓	χ	χ	✓	✓	✓	χ	χ	χ	χ	χ	L	N	This framework is primarily designed to group many of the earlier phases of digital forensics into preparation and identification, which comprises stages of collection, preservation and matching for evidence [16]
Integrated Digital Forensic Process Model (2013)[17]	Digital Forensics	χ	χ	✓	✓	✓	✓	✓	χ	✓	✓	✓	χ	χ	χ	χ	χ	χ	*L*	*N*	This process Model is a standardized model for the digital forensics method that helps the investigators follows a uniform approach in a digital forensic investigation based on different models.
An Examination of Digital Forensic Models	Digital Forensics	χ	χ	χ	✓	✓	χ	✓	χ	✓	✓	χ	✓	✓	χ	χ	χ	χ	*M*	*N*	The objective of this framework is to explore the development of digital forensic process models and construct specific forensic methodologies.
The Enhanced Digital Investigation Process Model	Digital Forensics	✓	✓	χ	✓	✓	✓	✓	χ	χ	✓	✓	✓	χ	χ	✓	✓	χ	*M*	*N*	The proposed framework aims to redefine the digital forensic process and progress.

Carrier and Spafford FW (2004)	Digital Forensics	✓	χ	✓	✓	✓	✓	✓	χ	χ	✓	✓	χ	✓	✓	✓	χ	χ	*M*	*N*	In this framework the techniques of digital investigation are categorized on the basis of something beyond the previous experiences and subjective preferences of the investigator
Integrated Digital Forensic Process Model (2013)	Digital Forensics	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	χ	χ	✓	χ	χ	*H*	*Y*	The main purpose of this framework is to propose a standardized model for digital forensics to help the investigators to follow a uniform approach in digital forensic research.
Data reduction and Data mining Framework for Digital Forensic Evidence (2013)	Digital Forensics	✓	χ	✓	✓	✓	✓	✓	✓	✓	✓	✓	χ	✓	✓	✓	✓	χ	*H*	*Y*	The goal of this framework is to reduce the volume of collected evidence to improve the review and investigation process.
Network Intrusion Forensic Analysis using Intrusion Detection System	Computer Forensics	χ	χ	χ	✓	✓	✓	χ	χ	χ	✓	χ	χ	✓	χ	χ	✓	✓	*H*	*N*	The author shows major challenges in computer forensics, intrusion detection system specialized model.
UML Modeling of Digital Forensics Process Models (DFPMs)	Digital Forensics	χ	χ	χ	✓	✓	✓	χ	χ	χ	χ	χ	χ	✓	χ	χ	χ	χ	*L*	*Y*	A lot of digital forensics process model has been successfully used in digital forensics, proposed model aimed to model some of the proposed process by using UML specifically the behavioral, use cases and activity diagrams.
Computer Forensics Investigation an Approach to Evidence in Cyberspace	Computer Forensics	χ	χ	χ	✓	✓	✓	χ	χ	χ	✓	χ	χ	✓	χ	χ	✓	✓	*H*	*Y*	This framework aims to define a new approach to solve and enhance the stage of computer forensics examination. The model meets Italian legislation and could probably be used in other countries as well.

Mapping Process of Digital Forensic Investigation Framework (2008)	Digital Forensics	χ	χ	χ	✓	✓	✓	✓	χ	χ	✓	χ	χ	✓	χ	χ	✓	✓	*M*	*Y*	The framework aims to produce the mapping process between activities and output from each phase of investigation framework, resulting in the creation of a new framework to optimize the entire Investigation process in digital forensics.
Common Process Model for Incident and Computer Forensics (2007)	Computer Forensics	χ	χ	χ	✓	✓	✓	χ	χ	χ	χ	χ	χ	χ	χ	χ	✓	χ	*L*	*N*	A new framework proposed for both Incident Response and Computer Forensics processes that combine their advantages in a flexible way: it allows for a management oriented approach in digital investigations while retaining the possibility of rigorous forensics investigation.
A Blockchain-based Process Provenance for Cloud Forensics (2017)	Cloud and Digital Forensics	χ	χ	χ	✓	✓	✓	✓	✓	χ	✓	✓	✓	✓	χ	χ	✓	χ	*M*	*Y*	The proposed framework uses blockchain technology in order to increase overall investigation efficiency and reliability.
An Integrated Lightweight Block chain Framework for Forensics Applications of Connected Vehicles (2018)	Internet of Vehicles and IOT Forensics	χ	χ	χ	✓	✓	χ	✓	✓	✓	✓	✓	χ	χ	χ	χ	χ	χ	*L*	*N*	The proposed framework uses block chain technology in order to improve safety and integrity of the collection stage.
Block chain based digital forensics (2017)	Digital Forensics	χ	χ	χ	✓	✓	χ	✓	✓	χ	✓	✓	✓	✓	χ	χ	χ	χ	*L*	*Y*	The proposed framework uses block chain technology in order to achieve integrity, authenticity, security and auditability for investigation process.

TABLE. II. Main Stages for the Framework which Proposed in Early Days of Digital Forensics

Framework Name	Stage One and it's sub-stages	Stage two and it's sub-stages	Stage three and it's sub-stages	Stage Four and it's sub-stages	Stage five and it's sub-stages	Stage Six and it's sub-stages	Comments
Lee FW (2001)	Recognize (Document and collect and preserve)	Incident recognition	Identify Sub-stage Classify Sub-stage Compare Sub stage	Individualize stage Evaluate sub-stage Interpret sub-stage	Reconstruction Stage Reporting Sub-stage Presentation Sub-stage		The proposed framework has three stages: identification, individualization and reduction. This framework can be considered as a framework of medium complexity as there are not many phases.
Casey FW (2004)	Incident recognition	Assessment Stage	Resource identification and seizer stage	Preservation Stage	Examination Stage Recovery harvesting Reduction classification	Analysis sub-stage Reporting sub-stage	The proposed framework contains all stages of the conventional framework with an addition to new stages such as assessment, incident recognition stages. This framework can be classified as a highly complex framework based on the addition of new stages to the standard stage and on numerous sub-stages in the examination stage.
Cohen FW (2009)	Identification Stage	Collection Stage	Transportation Stage	Storage Stage	Examination Stage Analysis Interpretation Attribution Reconstruction	Presentation Sub-Stage Destruction Sub-Stage	Proposed framework contains all stages of standard framework with addition of some stages such as the storage stage in place of preservation stage and the stage of destruction which added as a new stage to make sure that all collected evidence was deleted from the investigator side. Proposed frame work can be classified as a medium complexity and efficient framework.
Baryamureeba and Tushabe FW	Readiness Stage operational Readiness sub-stage Infrastructure readiness sub-stage	Deployment Stage Detection and notification Sub-Stage Investigation conformation submission	Trace back Stage Digital crime scene investigation sub-stage Authorization Sub-Stage	Dynamite Stage Physical crime scene investigation Sub-Stage Digital crime scene investigation Sub-Stage Reconstruction Sub-Stage Communication Sub-Stage	Review Stage		Proposed framework contain stages completely different from the stages of conventional frameworks, such as the readiness stage which refers to the preparation and pre-investigation stages, this stage contains a lot of sub-stages which add a thing of complexity to the framework. This framework can be classified as a highly complex framework because it contains a lot of new stages and sub-stages, for each of these stages.

TABLE. III. NEW FRAMEWORKS FOR DIGITAL FORENSICS DEPENDING ON EARLY FRAMEWORKS

Traditional framework	Novel frameworks
The proposed framework DFRWS 2001[5] consists of four main stages, preparation stage, identification stage, authorization and communication stage.	The proposed SRDFIM framework, 2011[23], consisting of set of stages, preparation stage, scene securing stage, screening stage recognition, scene documentation, communication shielding, evidence collection, preservation, screening, analysis and presentation stage.
IDIP framework 2003 [24], this framework consists of a set of phases starting from the crime scene preservation phase, Crime Scene Survey, Crime Scene Documentation, Crime Scene Search, Crime Scene Reconstruction.	The proposed CFFTPM 2006 framework [13] depends primarily on the IDIP framework and the CFFTPM framework consists of two main stages, the preparation stage and the analysis stage. The preparation stage consists of a set of sub-stages such as preparation, collection, and preservation of identification, and the second main stage is the analysis stage containing sub-phases such as examination, analysis and presentation.
An Integrated Digital Forensic Process Model 2013 [6] consisting of stages such as documentation stage, preparation stage, incident stage.	DFaaS Framework 2014 [25], phases of the DFaas Framework are dependent on the main stages of the IDFPM framework, starting from digital evidence collection and authentication, collection phase in this framework is different from the original framework because the evidence collected will be stored in central storage. After that the next stage in the proposed framework is the examination phase that was carried out using the current examination tools, then the results of the tools used through the examination phase are stored in the centralized database, then the reduction and analysis for the extracted information and then the presentation phase.

TABLE. IV. MAIN DIGITAL FORENSICS FRAMEWORK FOR DIFFERENT FIELDS OF TECHNOLOGY

Framework Name	Martini Framework 2012 [26]	Quick and Choo Framework 2014[27]	Perumal Framework 2015[28]
Main Stages	Evidence Source Identification. Evidence Collection. Evidence Preservation. Examination Information. Analysis Stage. Reporting Stage. Presentation Stage.	Commence. Preparation. Identify and Collection. Reduction by Reduce Data Collection. Review and Data Mining. Open and Close Source Data. Evidence analysis. Presentation. Complete	Plan. Evidence Identification. Examination. Analysis. Archive and Storage.
Technology name which can use the given framework	Cloud Computing, proposed frameworks specialized to solve the problem of investigation in cloud level.	For any digital forensics process that deals with a huge volume of information	Internet of Things specialized for IOT investigation process.
Main encouraged points	Huge volume of data, transferring evidence from remote location. Volatile data. No possibility to physically seizing all the servers in a cloud computing environment.	Slow analysis and examination process based on huge volume of information gather through collection stage	The increasing number of IoT devices connected will increase the quantity of evidence required by any investigative process.
The Idea from proposed Framework	Overcome set of issues that faces previous frameworks when applied in cloud computing environment such as volatile data gathering, collection of metadata that can help on investigation process	Reducing the amount of collected information by acquiring a subset of the data by utilizing data reduction and conduct intelligence analysis through data mining	Define a standard operating procedure for investigation of IOT de vices
The original framework on which the framework proposed is based	McKemmish 1999[9] and Kent et al. 2006 [10].	McKemmish’s (1999) [9] framework with the intelligence analysis cycle.	Triage model and 1-2-3 zone model for volatile based data preservation [12].

C. Main Challenges in Digital Forensics Process

According to research papers [29] [30], and [28], the main challenges of digital forensics can be classified into categories as listed below:

1) The diversity problem: Due to the fact that the digital forensics primarily depend on evidence collected from several digital device types, like computers, tablets, servers, camera and others, each of these devices has its own data format and can pose an important challenge during the analysis phase, because the evidence gathered is not standardized.

2) The efficiency of current digital forensics tools: Many of the digital forensic tools developed have been intended for finding at least one part of evidence, but they can be useful in other applications including standardizing different formats, compressing the collected evidence, extracting critical information and other tasks.

3) The volume of data collected from digital devices is huge, because of the growing number of digital devices used in our lives today. This huge volume of information causes difficulties at various stages of the digital forensic research process, such as the problem through the phase of evidence analysis, examination phase and other problems.

4) The complexity of format: This problem arises from data format collected from various digital devices requiring complex data reduction and review techniques.

5) The unified time lining issue: Based on gathering evidence from multiple digital devices, where multiple sources present different time zone references, timestamp interpretations, clock skew/drift issues, and the syntax aspects involved in generating a unified timeline.

6) Lack of training and resources: Any researchers manually inspect the need for a specific tool to utilize through the investigation process, and for this lack of training and resources, which refers to one of the major challenges faced by the digital forensic, specific training is required for these tools.

D. The Importance of Digital Forensics

In recent years, digital forensics has become important as the computer and cellular markets have grown. Based on the increased demand on smartphones, computers, and digital dependent through daily process, the market for malware or spyware has increased, digital forensics encompasses a variety of duties, such as the ability to recover different digital data, recover deleted data such as deleted messages from smartphones, deleted log files for different browsers, analyze and extract information and detect and remove different digital malware's or spy-wares. Malware or spyware refers to a program that allows for attackers to spy on user activities. Both malware and spyware are considered as a cyber-crime that can be extremely detrimental to you as an individual [31].

Adding the ability to applied digital forensics in computer resources will help in ensuring that the overall integrity and survivability of your network infrastructure, other implementing computer forensics can help the organization if you consider that computer forensic as a new basic element in what is known as a defense-in-depth (Defense in depth is designed on the principle that multiple layers of different types of protection from different vendors provide substantially better protection) approach to computer security and network security. Other important point of digital forensics is that DF can track where the user was or things started wrong before deleting it, while others can track down hackers even if the most important part of a digital hard drive is destroyed. When digital forensics are ignored or miss practice, essential evidence inadmissible legal evidence is entreated or collected. Others may escape new legislation mandating regularity compliance. The correct application of the forensics models may contribute to the prosecution of offenders. Digital forensics can provide feedback on improving current mechanisms for prevention to prevent a repetition of the event.

III. Main Branches of Digital Forensics

Digital forensics is used to investigate crimes, where a digital device is used either as a tool in enabling the crime or as a target of the crime. As illustrated in Fig. 1 the digital forensics consists of a set of branches. Many process models have been proposed for each of these branches, through the following section we will go deeply into most of these branches and focus on the cloud forensics branch.

The main branches of digital forensics are shown in Table V, with detailed information about each branch. Moreover, Table V shows the differences between the five branches of digital forensics depending on a set of criteria such as goal, type of collected information, Coverage digital devices, and main stages.

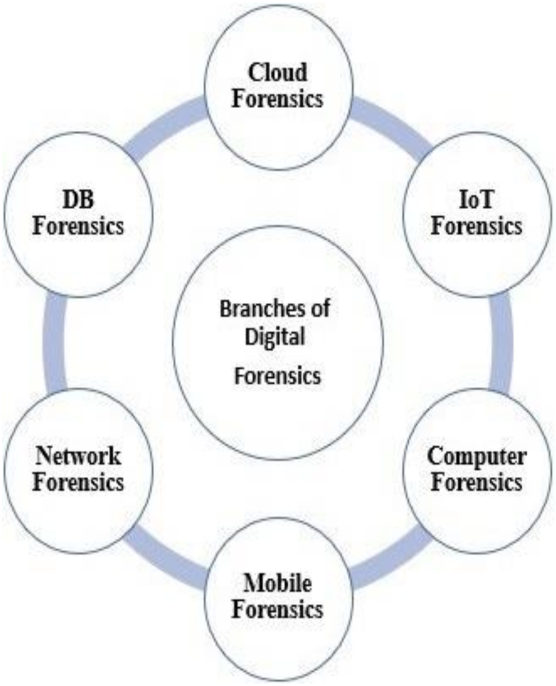


Fig. 1. Main Branches of Digital Forensics.

TABLE. V. COMPARISON BETWEEN MAIN TYPES OF DIGITAL FORENSICS BASED ON SPECIFIC PARAMETERS

Digital forensics branch name	Main Goals for DF Branch	Source of Evidence and type of information collected	Coverage digital devices	Main Stages of branch framework
Computer Forensics	The goal of computer forensics is to analysis information contained within and created with digital artifact; such as computer systems and electronic document.	Broad range of information start from log files such as inter net history, actual files stored in storage de vices, and static memory such as USB.	Traditional computers such as desktops, laptops, servers, tablets, etc.	Acquisition stage. Examination stage. Analysis stage Reporting Stage. Presentation Stage
Mobile Forensics	Recover digital evidence from a mobile device such as cellular phones, smartphones and mp3 players.	Communication information (SMS/ Emails), recovery of deleted data, contact numbers, photos in smartphones, notes, and other personal information	Mobile devices such as smartphones and mp3 players	Seizure stage. Acquisition stage. Examination and analysis stage.
Network Forensics	Monitoring, extracting and analysis of traffic on wired and wireless networks.	Routing tables, web browser history log, router logs, website pages, email attachments, VOIP data	Routers, internet applications, VOIP telephone, etc.	Identification stage. Preservation stage. Collection stage. Examination stage. Analysis stage. Presentation Stage.
DB Forensics	Study and analysis of databases and their metadata for incidents such as security attacks.	Database content, Metadata information, cached information which may locate in server RAM, database transactions, and queries.	Storage center, cash memory, servers RAM.	Identification Stage. Collection stage. Analysis Stage. Documentation Stage. And Presentation Stage [33].
IOT Forensics	Recovery of digital evidence form IoT devices such as sensors.	IOT applications, Smart home applications, sensor logs and information, and CSP log files	IOT devices such as sensor nodes, cars, smart applications.	Collection Stage. Examination Stage. Analysis Stage. Presentation Stage.
Cloud Forensics	Investigate the crimes that have occurred in cloud computing environments because it has many weaknesses, such as the nature of cloud computing, which can help to increase the level of crime	Users devices connected to the cloud computing environment, servers and storage center of cloud computing, cloud service provider	Laptops, desktops, smart phones, storage center devices and tablets	Preparation stage Identification stage Evidence collection Examination and analysis Presentation and Reporting

- Computer Forensics

Computer forensics can be defined as a discipline that combines legal and computer science elements to implement various stages of digital forensics on computer resources to explain the state of a digital artifact such as computer systems and electronic document [33]. The goal of computer forensics is to analysis of information contained within and created with digital artifact; such as computer systems and electronic document. Digital information required for the investigation process must be derived from a digital source in a timely manner, and critical information needed for the investigation process must be derived in a short time period [22].

- Mobile Forensics

Mobile forensics Recovers digital evidence from a mobile device such as cellular phones, smartphones and mp3 players. In recent years, mobile devices have been the booming technological trends along with Internet of Things, Cloud Computing and Big Data. Smartphones offer a range of features, allowing users to perform nearly every task done on computers. Fig. 2 history of the distribution of mobile devices (smartphones, tableting) compared to desktop computers can be found in the following graph. Smartphone replaces computers in almost every way since the majority of applications, from private use to business, from photos to online banking, are portable and more convenient to use. This means that for many investigations' smartphones have valuable information. They give recent chats, call logs, location data, photos, etc., to help the forensic investigator identify the person and learn about their recent work. They carry more personal data than a traditional PC in most cases.

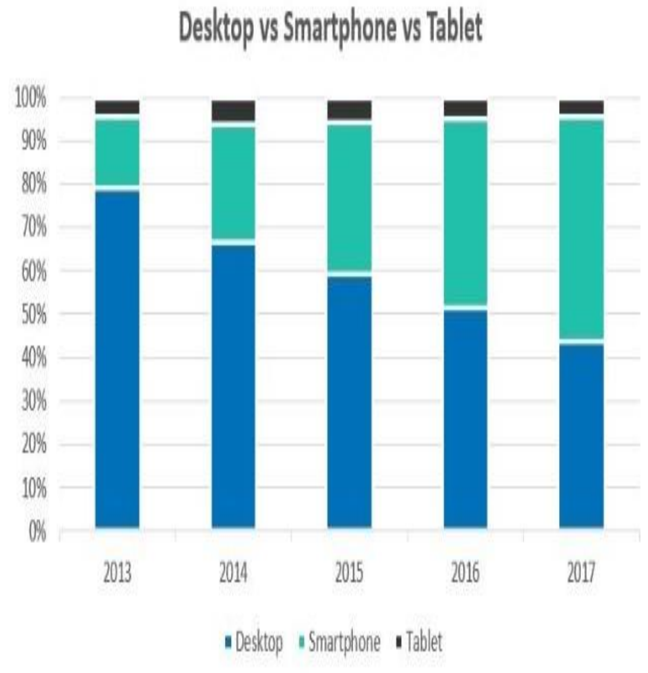


Fig. 2. History of the Distribution of Mobile Devices (SmartPhones, Tableting) Compared to Desktop Computers.

Many mobile forensic frameworks have been proposed, such as the framework proposed by Elhadje [34], the author motivated by the lack of mobile forensic frameworks to analyze and map data collected from various sources, including calls, geographical locations, multimedia, the proposed framework and others could contribute to a novel and potentially market-leading forensic mobile framework. Another framework proposed by M.Petraityte on the field of mobile forensics [35], the main objective of the proposed framework is to assist forensic investigators by potentially shortening the time spent investigating possible infringement scenarios. And many other frameworks and tools proposed with some update based on the nature of mobile devices in the field of mobile forensics with the same stages of digital forensics.

- Network Forensics

Network forensics is an extension of digital forensics, which can be defined as a network traffic investigation process that provides a means of monitoring various cyber-crimes by analyzing and tracking the evidence gathered from the network, and highlights the detection and prevention of different network attacks. Network forensics analysis tools can provide many functions such as network forensics and security investigation, data integrity from multiple sources, prediction of future targets for attacks, network traffic analysis, and recording various types of traffic analysis based on user needs, and many other functions can be provided by network forensics tools. The network forensics process can be applied on different network layers as follows [36][37]:

- Network forensics on Ethernet layer: The network forensics process can be applied on the Ethernet layer by eavesdropping bit streams with tools, called sniffing tools or monitoring tools. Famous tool for doing this is Wire-shark and d tcpdump, where tcpdump works primarily on Unix operating systems, which can assist the investigator in gathering various evidence on the Ethernet layer.

- Network forensics on network layer: Through network layer the internet protocol (IP) is responsible for delivery of packets generated by TCP through the network by adding both source address and destination address. The packets that sent from source node to destination node must go through set of routers, each router contains a routing table. The Routing tables are one of the sources for evidence in network forensics process, because it can help to track down the attackers by reverse the sending route and find the computer the packet came from. Other evidence resources in network forensics refer to network device logs that record traffic information. To reconstruct the attack scenario, multiple logs recorded from different network device can be correlated together. Network devices have limited storage capacity and network administrator can configure the devices to send logs to a server and store them for a period of time. The Network forensics have received a great importance due to the following facts:

1) Due to a large number of attacks through the network, like the DDOS attack on the social network and push attacks, different organization is mainly concerned about their network and data transfers through the network locally or publicly. For this, it is necessary to trace out the criminals, and collecting legal evidences is required through the trace operation to present them in the court. Based on this it is necessary to have forensics principles for network environment to collect evidence that can be used in the court of law.

2) For any investigation process it is necessary to use network forensics to analyze the historical network data in order to investigate security attacks by reconstructing sequence of security attacks.

3) Network forensics can do other than investigation on crimes happened on the network level. Network forensics can be used to address network issues of business-critical systems.

4) Network forensics is important to get trustworthy of users, and the ability to safeguard their interest. Network forensics can provide monitor and analyze their network traffic to detect malicious events and take actions for attack as quick as possible.

The network forensics branch faces a set of challenges such as data and traffic-related challenges. The graph below shows the main challenges faced by the network forensics, and proposed solutions [38]. The challenges faced by the forensic network have been identified in Fig. 3. various frameworks and solutions have been proposed in the field of network forensics, these solutions can be classified as follow:

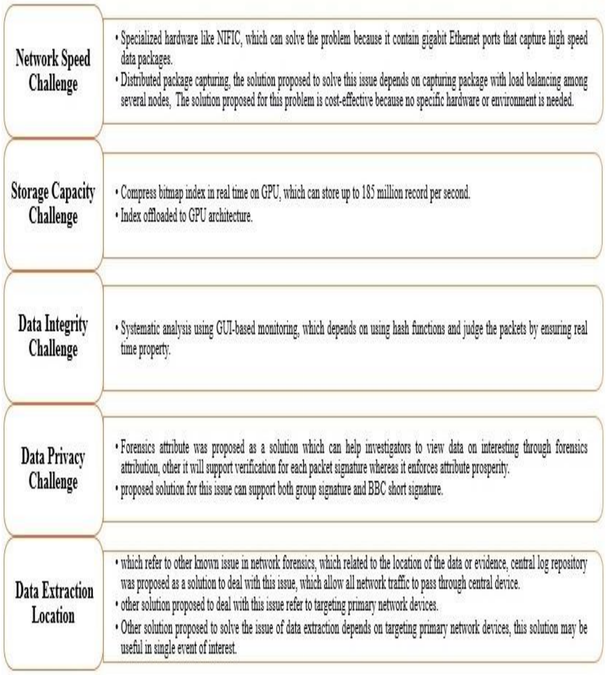
- Distributed Network Frameworks and Solutions. A lot of frameworks and solution proposed in network forensics can fall under this type of network forensics such as the framework proposed by Shanmugasundaram et al. through [39], the author proposed a distributed network logging mechanism over wide area networks. [U+0650] Another distributed frameworks have been proposed through [40], [41]. And The framework of Mandia and Procise [42], which adding the two-way incident response stage between detection and presentation stages.

- Dynamic Network forensics frameworks, which used in large scale environment and depends on storing collected evidence in distributed DB to achieve level of security various frameworks fall under this category of network forensics frameworks such as the framework proposed in 2007 by Wang et al [43]. which proposed the model based on the artificial resistance theory and multi agent theory and the model proposed through [15], Kohn Framework [6], follows standard steps in digital forensics investigation processes, And the framework proposed by Liu in [44], which depends on using a logic-based network forensic process model using PROLOG to analyze the evidence collected and delete all unrelated data,

- Soft computing-based network forensic frameworks. Specified for an unsecured environment and environment that contains many attacks because this category of frameworks deals with the analysis of data collected and the classification of related attacks. This category of network forensics frameworks involves various frameworks such as [45], [46], and [47].

- Graphic Based Network Forensics Frameworks. Various framework were fall under this category such as the framework which proposed in 2008, Wang and Daniels. Proposed framework is graph-based approach for network forensics analysis [40] and [48].

- Internet of Things (IOT) Forensics v Internet of Things can be viewed as an information system made up of things, networks, data, and services. Such things may be wireless sensors, traditional computers, cameras, home appliances, tablets, smartphones, vehicles, humans, etc. that are connected over a network which can be wired or wireless. These things may gather, process, and upload a huge amount of data to the internet and used to initiate service. The architecture of IoT combines different zones such as perception zone, fog zone and cloud zone, where each zone can be a source of evidence in IoT forensics, such as evidence that can be selected from smart IoT devices, sensor nodes, firewalls, routers, etc. IoT forensics depends mainly on the main stages of digital forensics investigation such as the collection, examination, analysis and presentation of digital evidence with difference in some points such as source of evidence. IoT forensics could handle any possible format of data evidence, in traditional forensics it should handle electronic documents or standard format [49] [50]. IoT forensics process faces many challenges such as data location as many of the IOT devices are distributed in various locations that are out of user control, which can affect the investigation process. Moreover, IoT forensics process faces another challenges such as the limitation of the lifespan of digital media and the limitation of storage devices [51]. Table VI appears the famous proposed IoT forensics frameworks.



Network Speed Challenge	• Specialized hardware like NIFIC, which can solve the problem because it contain gigabit Ethernet ports that capture high speed data packages. • Distributed package capturing, the solution proposed to solve this issue depends on capturing package with load balancing among several nodes, The solution proposed for this problem is cost-effective because no specific hardware or environment is needed.
Storage Capacity Challenge	• Compress bitmap index in real time on GPU, which can store up to 185 million record per second. • Index offloaded to GPU architecture.
Data Integrity Challenge	• Systematic analysis using GUI-based monitoring, which depends on using hash functions and judge the packets by ensuring real time property.
Data Privacy Challenge	• Forensics attribute was proposed as a solution which can help investigators to view data on interesting through forensics attribution, other it will support verification for each packet signature whereas it enforces attribute prosperity. • proposed solution for this issue can support both group signature and BBC short signature.
Data Extraction Location	• which refer to other known issue in network forensics, which related to the location of the data or evidence, central log repository was proposed as a solution to deal with this issue, which allow all network traffic to pass through central device. • other solution proposed to deal with this issue refer to targeting primary network devices. • Other solution proposed to solve the issue of data extraction depends on targeting primary network devices, this solution may be useful in single event of interest.

Fig. 3. Challenges Face network Forensics and Proposed Solutions for each Challenge.

TABLE. VI. MAIN FRAMEWORKS AND SOLUTION THAT PROPOSED IN IOT BRANCH OF DIGITAL FORENSICS

Framework Name	Initialization Stage	Identification Stage	Evidence Collection Stage	Reduction Stage	Preservation Stage	Examination Stage	Analysis Stage	Sharing With Another Investigators	Review Stage	Documentation Stage	Presentation Stage	Reporting Stage	Complexity Level	Comments And Limitation	Main Contribution
Generic Digital Forensics For IOT [52]	✓	✓	✓	χ	✓	✓	✓	✓	χ	χ	✓	✓	(*H*)	The proposed framework contains the basic stages of the digital forensics process, but it does not contain any strategy for feedback and evaluation, and the proposed framework does not concern privacy and integrity.	The forensics framework has been suggested as there is no framework available for investigating digital crimes that can occur in a smart environment. The importance of the proposed framework stems from the various safety principles proposed by the ISO standard
IoT forensics framework for smart environment [53]	✓	✓	✓	χ	✓	✓	✓	χ	χ	χ	✓	χ	(L)	The proposed framework is a privacy conscious framework that takes into account a set of privacy principles that can improve data privacy but, at the same time, is not suitable enough for IoT devices with limited resources.	The main goal for the proposed framework is to introduce a new lightweight version of the IoT forensics framework, that can be applied to investigate crimes that have occurred in the IoT environment to be suitable for the nature of the IoT devices.
Privacy aware IoT forensics process model [54]	✓	✓	✓	χ	✓	✓	✓	✓	✓	✓	✓	✓	(H)	The proposed framework is a privacy conscious framework that takes into account a set of privacy principles that can improve data privacy but, at the same time, is not suitable enough for IoT devices with limited resources.	Through this research paper, novel privacy conscious IoT forensics process model was proposed. The idea here is to achieve different principles proposed by ISO / IEC 29100:2011. The proposed framework consists of all basic stages of any forensic process model with the addition of new stages such as review stage, initialization stage, and feedback stage

IoT-Forensics Meets Privacy Towards Cooperative Digital Investigations [55]	✓	✓	✓	χ	✓	✓	✓	✓	χ	✓	✓	χ	(H)	The author has suggested an improved Model for an IoT environment investigation process that takes into consideration many of the concepts of privacy and security.	The idea from proposed model is to present a digital witness approach with methodology that enable citizen to share its sensitive information with investigators based on using PROFIT methodology.
Application Specific Digital Forensics Investigative Model in IoT [56]	χ	✓	✓	χ	✓	✓	✓	χ	χ	✓	✓	✓	(L)	The proposed model consists of basic stages of the forensic process, It does not take into consideration any principles such as privacy, integrity, and other security principles. .	The idea from this research paper is to introduce an application for a specific digital forensic investigation model, which will be applicable with any IoT related forensics investigation. Artifact of forensic importance in three highly adopted IoT applications scenario, smart home, smart city and wearable.

- Cloud Computing Forensics

Cloud computing can be defined as an internet-based computing paradigm in which a huge number of computer system resources such as computing power and storage are networked to enable users to remotely access and use these resources via the internet. Cloud computing can be categorized according to service model into three major types:

1) Infrastructure as a Service (IaaS) which delivers basic computer infrastructure as a service such as network capability and storage space.

2) Platform as a Service (PaaS) which refers to delivery of an entire computing platform and solution stack as a service.

3) Software as a Service (SaaS) which refers to provide a software which hosted centrally and can be accessed by any user using thin client, this not need to purchase and install.

A cloud forensics can be defined as a mixture of traditional computer forensics, digital forensics, and network forensics [57][58], other definition for cloud computing forensics was defined by [59] which define it as an application of digital forensics in cloud computing environments.

Other researchers, such as in [60], define cloud forensics as a part of network forensics that uses a fresh cloud-friendly technique to follow the primary stages of network forensics.

The suggested cloud forensics frameworks in [57] [59] show a set significance points of cloud forensic, these points and their evaluation results are as follows:

1) Cloud forensics is important component for cloud security.

2) There will be a lack of awareness until a major critical incident happens.

3) Cloud forensics is important to get the trust of the users who need to use cloud resources.

4) Cloud forensics needs more funding and investment in the RD than it has got at the moment.

The cloud forensics faces many challenges, here we exam ine set of challenges in cloud computing forensics as follows:

1) Unification of logs format: The cloud computing consists of a huge number of servers, where each server in cloud environment has its log format and this will hamper the investigation process as the evidence gathered will be in more than one format, this point makes the investigation more complex. Synchronization and time stamp, because of a large number of server's participants in a cloud system and the distributed locations for these servers, each of locations have a specific time zone. This can cause a problem through the investigation process in a cloud environment. One of solutions that suggested to solve the time zone problem is suggested by [61], which suggest a specific time system to be used on all entities of the cloud. This can achieve benefit of having a logical time pattern.

2) Missing terms and conditions in service level agreement (SLA) regarding investigations, where service level agreement is the main points and conditions between the user and the cloud service provider. These points should include important terms regarding cloud forensics investigation. Ruan et a1. [60] SLA should include: service provided, technique supported, access granted by CSP to the customer, security issue in multi-jurisdictional environment in terms of legal regulation, privacy policy and customer data.

3) Lack of forensics expertise, especially on the level of cloud computing.

4) Decrease access to forensic data and control over forensics data at all level from customer side. Itemizem Single point of failure in cloud forensics investigation process, because evidence from different servers on the cloud must be stored in central server for investigation process.

5) Lack of international collaboration and legislative mechanism in cross-nation data access and exchange. Specially because cloud forensics depends on collecting evidence from servers located at different countries

6) Integrity and stability, cloud forensics depends on client server communication. Evidence transferred over public network between investigator machine and cloud storage device, integrity of evidence is very important point through investigation process. Many of solutions suggested to enhance the integrity level of cloud forensics investigation process such as in [62], which suggest a digital signature for all collected evidences through evidence collection stage and then check this signature on the other side and when start examination stage. Other solution suggested for this problem by Hegarty [63] which implement a specific framework for digital signature detection that enables forensics analysis of storage platform.

The differences between the investigation of crimes committed in the cloud computing environment and those committed in conventional digital devices are as follows:

1) Conventional digital forensics cannot be used to investigate the incident in cloud computing on the basis of various factors such as the distributed nature of cloud computing [64][65], large resources rather than limited resources in local devices, a large number of data centers located in multiple locations, the presence of third parties and many other factors.

2) Another important factor makes cloud forensics unfriendly and different from digital forensics is that cloud forensics cannot confiscate the suspicious computers and have direct and physical access to the resources that may contain the evidence. This is because all of these evidences are far-reaching and can be found elsewhere.

3) The privacy of user's data, due to the fact that cloud computing contains information from various users, the investigator needs to access total cloud-level data in the data center, user-related information and other users for the purpose of extracting and examining evidence. This process is different and makes cloud forensics uncomfortable from other branches of digital forensics.

4) The nature of cloud computing is complicated, which consist from large number of resources and collect a huge volume of evidence. Based on that a parallel implementation using distributed system is required to enhance the performance of investigation process [66] [67] [50].

5) In general, cloud computing depends on the cloud service provider "CSP" in collecting cloud data, CSP is not a legal investigator, and the trust of cloud service providers is yet another major challenge in implementing digital forensics in a cloud environment that is different to the traditional process of digital forensics investigation.

6) The custody chain differs from that of digital and cloud computing forensics, the custody chain clearly deduces the way evidence has been found, gathered, analyzed and maintained to be submitted to the Court in a manner permissible [11]. In conventional digital forensics, it is trivial to access the history, location, and main resources of the computer. On cloud computing, the location of sources that may contain evidence is hard to identify, because each cloud server is located within a given geographical area, and the time zones for various cloud servers differ.

7) Cloud computing is a system for multiple tenants, while traditional computers are a single owner. This point has a major problem during the investigation, because different users or virtual machines share the same physical resources, and can cause a problem because the investigator has to provide a court with the information collected concerning the suspected user, as the alleged suspect may say that the information contained in the evidence contains data for other users. Another problem here is the privacy of the information of other users, as the investigator can access various user resources.

8) The Cloud Process is limited by Cloud Service Provider and the User-Cloud Provider contract is important because the User-CSP contract includes the Service Level Agreement between the two sides. You could be in an extremely bad situation if this agreement doesn't show at what level your service provider is obliged to provide forensic information, and also how soon it is necessary to do so.

Several investigation frameworks were offered for crimes committed in cloud computing environments, such as the model that was proposed in [68]. A cloud forensics model includes key stages in all models of digital forensics processes such as identification stages, collection stages, preservation stages, examination stages, analyses, and presentations stages. The author develops a proposed model as a service (FPaaS) using cloud-based business process using cloud-based business execution language (BPEL), which combines the four main stages into a single service called (FPaaS). Another model proposed by Shan and Malik [69], which includes three major stages, the identification stage, data collection and preservation stage, and the analytical and presenting stage, is another process model proposed for the cloud computing environment, and, through the suggested model, defines the key challenges which can face each phase within the cloud computer environment.

Martini and choo (2012) [26] suggest a new forensic cloud framework; a proposed model utilizing the main stages of the conventional digital forensics model with enhancement; a principal enhanced point in a model is the iterative stage after the preservation stage, where evidence gathered through the phased evidence is not enough; then the process retraces to the stage of identification. The combination of the two stages of identification, collection and preservation into a unique step is another enhanced feature in the Martini and Cho Framework because evidence can be eliminated or modified at any time in a cloud environment is highly likely to be science and the cloud environment

An Open cloud forensics process model (OCF) was proposed through [70] by Shams Zawoad. This model focuses on the main roles of the cloud service provider, and how the provider effects on the forensics process on the cloud level, and suggest a role of the cloud service provider to support reliable digital forensics in the cloud. Another contribution for this model is to extend the definition of digital forensics investigation process to support reliable digital forensics in the cloud, other contribution for this model that it presents a new architecture which called forensics-aware cloud architecture. The proposed model consists from the basic stage of digital forensics process mode with the addition a verification stage as a last stage of process model.

Table VII shows the key frameworks suggested in a cloud computing environment, the table contains information about each of the key frameworks such as the name of the framework, the contribution for each of the key frame and the goal from proposing each framework, and main stages of the proposed framework by suggesting a set of stage and specific column called other stages if the framework contain stages more than what suggested, and the weak points and drawbacks if exist for each of the listed frameworks. Other process models have been proposed to solve the issues that face the process model of digital forensics when applied in a cloud computing environment. The main challenges facing cloud computing forensics and the solutions proposed to solve them are as follows:

1) The challenges faced by the identification stage are shown in Fig. 4. The first challenge is the access to evidence in log files, which means that it is difficult to access such files in cloud environment due to the fact that there are different formats of log files. Several solutions have proposed to the above challenges in [73], [74], [75], [76], Table VIII shows the main solutions that proposed to deal with the challenges faces the identification stage in the investigation process, Table VIII contain solution name, author name, solution contribution, and solution drawbacks.

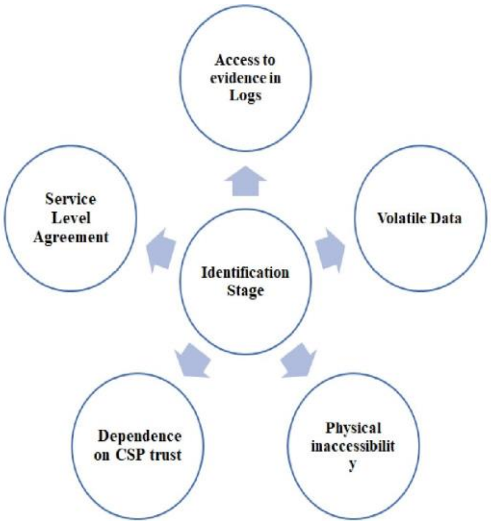


Fig. 4. Major Challenges in Cloud Computing Forensics through the Identification Stage.

The second challenge faced by the identification and collection stage is the volatile data problem. Several solutions have been propose to solve this challenge such solution introduced by Grispos through [77], Grisopos in [77] suggested a new concept to overcome the volatile problem (unstable data) by using a particular strategy that allows the investigators to collect data that may be lost whenever possible. Another solution to volatile data problem proposed by Damshenas and Brik [61], [1] which suggests a frequent data synchronization between virtual machines and persistent storage.

The third challenge confronting the identification and collection stage is the service level agreement “SLA”, where “SLA” relates to the primary contract between cloud participant and cloud service provider. This arrangement must include various points that can assist the investigator. The solution to this challenge has been proposed by many researchers as in, [15]. The suggested solution in, [15] comprises a set of conditions to regulate the agreement between the CSP and the participant. Another solution to this challenge proposed by Damshenas [16], And Baset [36]. Their solution comprises a guideline which explains how the SLA should be implemented. In addition to the above solutions, Brik [37] and Haeberlen [78] proposed another solution which suggests a third-party confidence to audit the safety measure given by the CSP.

The fourth challenge in the identification and collection stage is physical inaccessibility due to the nature of cloud computing environment which is geographically spread by the hardware devices as discussed in [62] [79].

The fifth challenge in the identification and collection stage is the dependence on cloud service provider which affects the trust between both CSP and users of the cloud. The absence of transparency and confidence between CSPs and clients is a problem that has been dealt with by Haeberlen [80], which is a primitive fundamental called AUDIT that could be provided by an accountant. Other model called trust cloud framework have been proposed by Ko et al. [81], which consists of five layers of accountability: system, data, workflow, policies, and laws and regulations layers. To increase accountability detective approaches used rather than preventive.

1) The Challenges faced by the analysis stage One of the main challenges faced by the analysis stage is using the encryption algorithms for examination and evaluation stage. In [82] the author proposed a hierarchical attribute- set- based solution to overcome the above challenge. The proposed solution is applied to accomplish fine- grained access control in cloud computing which achieves scalability and flexibility more than confidentiality and authenticity. Another challenge in the achievement of confidentiality and authenticity is through the examination stage. The solution to deal with this challenge was proposed by Prabha N et al. through [83], which presents an encryption technique for query processing on a cloud to protect confidentiality and authentication.

TABLE. VII. MAIN FRAMEWORKS THAT SUGGESTED FOR A CLOUD COMPUTING FORENSICS

FW. Name	Contribution	Identification stage	Evidence Collection	Preservation Stage	Examination and Analysis Stage	Presentation and Reporting Stage	Other Stage	Drawbacks
An integrated digital forensics framework for cloud computing [71]	Define the difference between evidence collection stage and preservation stage in investigation process	✓	✓	✓	✓	✓	x	The main idea of proposed framework is to show that preservation stage is different from collection stage. That's right but a lot of previous work shows that both stages must be at same level to prevent the evidence that collected from any change and update. No any action taken on the other stages of forensics process to be suitable with cloud environment.
Forensics investigation process [57]	A new framework for cloud computing environment with basic change on the main stages of original digital forensic phases	✓	✓	✓	✓	✓	✓	The framework does not take any action for other stages rather than identification stage and collection. The stages of preservation, analysis, examination in digital forensics needs a lot of updates to be suitable for cloud forensics. The author does not take any action in proposed framework to enhance the security and privacy of cloud user's data.
An Open Cloud Forensics model [56]	Main contribution is to run preservation stage in parallel with all other stages of digital forensics stage. Other update added by proposed framework is a combination between examination stage and analysis stage into a new stage called organization stage.	✓	✓	✓	x	x	✓	The author through proposed idea run the preservation stage in parallel with each of original stages of digital forensics, this step will do an overhead on the forensics process because actually it just required in the step of evidence collection, not at each step of forensics process. The author does not add any action to achieve the main principles of security and privacy for user's data on the cloud.
Adams Process model [72]	A new model specified for acquisition stage on cloud computing, define the documents and resources that may contain the evidences that related to the crime.	✓	x	x	x	x	✓	The author proposed an enhanced way for acquisition stage. But no action was taken to enhance other stages to be suitable for cloud computing. The process of enhancement was not completed because the author proposes an enhancement on the level of identification. To complete the process the author must add some enhancement on the level of evidence collection. A lot of principles in the level of security and privacy must be taken into account when trying to design an investigation model which deal cloud computing.
Shah Framework [1]	A new Framework proposed for evidence collection on cloud computing environment, proposed model consist from set of layers, each layer contain set of stages of forensics process model	✓	✓	✓	✓	✓	x	Main drawback in proposed model that this model does not add any specific action on any of digital forensics stages to be suitable with cloud computing environment. The author says that proposed model will deal with the dynamic nature of cloud computing by combining set of stages at same layer, dynamic nature for cloud forensics needs an action to achieve highly level of security on evidence transferring between server and investigator side, and other an action is required to achieve a level of privacy on user's data. The author in his research paper focused on an important issue which faces the process on the investigation in cloud computing this problem is dependent on CSP, no action has been taken by the model to deal with this problem.

TABLE. VIII. MAIN SOLUTIONS THAT PROPOSED TO SOLVE THE ISSUE RELATED TO LOGGING IN IDENTIFICATION STAGE

Solution Name	Author name and year of publish	Contribution for the solution	IaaS	PaaS	SaaS	Drawbacks
Secure logging-as-a-service for cloud forensics [73]	Zawoad S, Dutta AK, Hasan R. SecLaas (2011)	Introduce a secure logging as a service which allow the CSP to store virtual machine logs and provide access to forensics investigators while preserve the confidentiality for the cloud users.	✓	✓	✓	Proposed solution still depends on cloud service provider in the proposed solution, where CSP may not have a level of trust. The other comment is the location of storing VM logs. And how these logs will be stored as clear text or as an encrypted format.
log-based approach to make digital forensics easier on cloud computing [74]	Sang T. A (2013)	Propose a log-based model, the idea of proposed model is to keep another log locally and synchronously. So, it can be used to check the activity on SaaS without interference from CSP. The main goal for this solution is to decrease the interference on cloud service provider on the investigation process. Other goal for proposed solution is to reduce the complexity of the forensics process in cloud computing.	χ	✓	✓	Proposed idea depends on using a SaaS in the cloud side, the user must initiate a request asking for log activity, SaaS will receive the request and then process the request and send it back the response to the user. The process of request and receive a log activity will constitute a great load on the network and will be not secure enough, because there are many of attack in the middle can do sniff and analysis for the traffic because log information must be sent when a small change on the log file happened
Digital forensic readiness in the cloud[75]	Trenwith PM, Venter HS (2013)	The author proposed a model that consider centralized logging for all activities of all participants of the cloud as solution to provide an efficient forensics strategy. Proposed model will enhance and quicken the investigation process.	✓	✓	✓	Main drawback of proposed solution is the centralized location which can have a central point faller, other it is not secure enough when all logs and activities stored at central location. Other main drawback refers to load balancing and heavily traffic on central point
Cloud application logging for forensics [76]	Marty R (2011)	The goal from proposed idea is to provide a lot of information about each record on the log such as when the log happened or record who triggered the event and why it happened, based on that the information that needs to be present at each record will be limited	χ	χ	✓	Proposed idea suggests a log management SaaS to define what is the main field needed at log file, but at the same time it does not provide any solution about logging network usage, file metadata, and other evidence which are important for investigation process in PaaS and IaaS.

2) The multi-tenancy challenge faced by the all stages in investigation process in cloud computing environment. This challenge means that the investigators can access all data of users in the cloud which leads to violate the privacy of users. The solution to this problem was proposed by Aydin M et al. [84]. The solution uses a third party to check and evaluate the data collected by the investigators. Another solution to solve the multi-tenant problem has been proposed by Martini through [26], which proposed a solution depends on upholding the confidentiality and integrity for the evidence that used through the investigation process because the cloud computing environment nature is multi-tenant.

3) The data gathering challenge faced by the acquisition stage: The data is distributed among different servers in cloud computing environment which decrease the performance of the investigation process. The solution to this problem was proposed by Adams in [72] which presents a new cloud forensics model that consists of initial planning stage, on-site survey stage and the acquisition of electronic data.

4) Solution suggested to solve the problem which faces the acquisition stage, Adams in [72] introduces a new cloud forensic model specified for cloud computing environment. The proposed framework specified for evidence acquisition. Not take any action for other stages of forensics process. Proposed model consists from three main stages as follow (a) initial planning stage, which specified for defining and determining all related documents that associated with the investigation, (b) the on-site survey stage, which define all knowledge relating to the location, size, and format of the device that may hold information that can help the investigation process, (c) the acquisition of electronic data which include the process of gathering data related to the crime and store these data. Main drawback for propose that focus on defining and acquiring digital data but does not deal with the stages of analysis and presentation.

IV. CONCLUSION AND FUTURE WORK

Various frameworks and solutions to deal with the process of digital forensics have been proposed. Some of these frameworks and solutions have been proposed as a general framework for digital forensics, while others have been proposed for a particular class of digital forensics such as IoT forensics, network forensics, and other classes. We discussed the main classes of digital forensics through this survey, the major frameworks suggested for each class, the principal steps of each framework, the problems to be solved within the framework, and the major disadvantages of each.

This survey primarily presents and compare between different frameworks suggested for the cloud computing investigation and all other digital forensics classes. We can conclude that most of the frameworks included in our survey focus on solving a specific issue related to CSP, logging issue, or other similar issues. Up to our knowledge we have not found any framework that takes into consideration the security and privacy issues that are becoming very important issues in cloud computing, especially when dealing with remote servers and cloud participants documentation. In this context, we have recommended a solution to improve many key issues such as security, accuracy, performance and privacy.

References

[1] L. Daniel, Digital forensics for legal professionals: understanding digital evidence from the warrant to the courtroom. Elsevier, 2011.

[2] C. Hargreaves and J. Patterson, "An automated timeline reconstruction approach for digital forensic investigations," Digital Investigation, vol. 9, pp. S69–S79, 2012.

[3] "USLegal-Definitions,"http://definitions.uslegal.com/d/digitalevidence, 2019.

[4] N. T. A. Recipes-A, J. McCaffrey, V. T. Patch, S. Manzuik, P. Chandra, M. Messier, J. Viega, O. D. Wiley, P. Elst, Y. T. Apress et al., "of the book author publisher."

[5] M. Reith, C. Carr, and G. Gunsch, "An examination of digital forensic models international journal of digital evidence," 2002.

[6] M. D. Kohn, M. M. Eloff, and J. H. Eloff, "Integrated digital forensic process model," Computers & Security, vol. 38, pp. 103–115, 2013.

[7] F. Servida and E. Casey, "Iot forensic challenges and opportunities for digital traces," Digital Investigation, vol. 28, pp. S22–S29, 2019.

[8] E. Casey, Digital evidence and computer crime: Forensic science, computers, and the internet. Academic press, 2011.

[9] R. McKemmish, What is forensic computing? Australian Institute of Criminology Canberra, 1999.

[10] K. Kent, S. Chevalier, T. Grance, and H. Dang, "Guide to integrating forensic techniques into incident response," NIST Special Publication, vol. 10, no. 14, pp. 800–86, 2006.

[11] J. R. Vacca, Computer Forensics: Computer Crime Scene Investigation (Networking Series), (Networking Series). Charles River Media, Inc., 2005.

[12] E. S. Pilli, R. C. Joshi, and R. Niyogi, "Network forensic frameworks: Survey and research challenges," digital investigation, vol. 7, no. 1-2, pp. 14–27, 2010.

[13] M. Köhn, M. S. Olivier, and J. H. Eloff, "Framework for a digital forensic investigation." in ISSA, 2006, pp. 1–7.

[14] G. Palmer, "A road map for digital forensics research-report from the first digital forensics research workshop (dfrws)," Utica, New York, 2001.

[15] B. Hitchcock, N.-A. Le-Khac, and M. Scanlon, "Tiered forensic methodology model for digital field triage by non-digital evidence specialists," Digital investigation, vol. 16, pp. S75–S85, 2016.

[16] M. Rogers, "Dcsa: Applied digital crime scene analysis," Tipton & Krause, 2006.

[17] S. Von Solms, C. Louwrens, C. Reekie, and T. Grobler, "A control framework for digital forensics," in IFIP International Conference on Digital Forensics. Springer, 2006, pp. 343–355.

[18] H. C. Lee, T. Palmbach, and M. T. Miller, Henry Lee's crime scene handbook. Academic Press, 2001.

[19] S. Ó. Ciardhuáin, "An extended model of cybercrime investigations," International Journal of Digital Evidence, vol. 3, no. 1, pp. 1–22, 2004.

[20] B. Carrier and E. H. Spafford, "An event-based digital forensic investigation framework," in Digital forensic research workshop, 2004, pp. 11–13.

[21] V. Baryamureeba and F. Tushabe, "The enhanced digital investigation process model," in Proceedings of the Fourth Digital Forensic Research Workshop, 2004, pp. 1–9.

[22] M. K. Rogers, J. Goldman, R. Mislan, T. Wedge, and S. Debrota, "Computer forensics field triage process model," Journal of Digital Forensics, Security and Law, vol. 1, no. 2, p. 2, 2006.

[23] S. A. Ali, S. Memon, and F. Sahito, "Challenges and solutions in cloud forensics," in Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing. ACM, 2018, pp. 6–10.

[24] M. Mabey, A. Doupé, Z. Zhao, and G.-J. Ahn, "Challenges, opportunities and a framework for web environment forensics," in IFIP International Conference on Digital Forensics. Springer, 2018, pp. 11–33.

[25] S. Raghavan, "Digital forensic research: current state of the art," CSI Transactions on ICT, vol. 1, no. 1, pp. 91–114, 2013.

[26] J. Yadav, "The impact of digital forensics in future," 03 2017.

[27] A. Agarwal, M. Gupta, S. Gupta, and S. C. Gupta, "Systematic digital forensic investigation model," International Journal of Computer Science and Security (IJCSS), vol. 5, no. 1, pp. 118–131, 2011.

[28] B. Carrier, E. H. Spafford et al., "Getting physical with the digital investigation process," International Journal of digital evidence, vol. 2, no. 2, pp. 1–20, 2003.

[29] R. Van Baar, H. Van Beek, and E. Van Eijk, "Digital forensics as a service: A game changer," Digital Investigation, vol. 11, pp. S54–S62, 2014.

[30] B. Martini and K.-K. R. Choo, "An integrated conceptual digital forensic framework for cloud computing," Digital Investigation, vol. 9, no. 2, pp. 71–80, 2012.

[31] D. Quick and K.-K. R. Choo, "Data reduction and data mining framework for digital forensic evidence: storage, intelligence, review and archive," Trends & Issues in Crime and Criminal Justice, vol. 480, pp. 1–11, 2014.

[32] A. Yasinsac, R. F. Erbacher, D. G. Marks, M. M. Pollitt, and P. M. Sommer, "Computer forensics education," IEEE Security & Privacy, vol. 99, no. 4, pp. 15–23, 2003.

[33] A. Al-Dhaqm, S. Razak, S. H. Othman, K.-K. R. Choo, W. B. Glisson, A. Ali, and M. Abrar, "Cdbfip: Common database forensic investigation processes for internet of things," IEEE Access, vol. 5, pp. 24 401– 24 416, 2017.

[34] E. Benkhelifa, B. E. Thomas, Y. Jararweh et al., "Framework for mobile devices analysis," Procedia Computer Science, vol. 83, pp. 1188–1193, 2016.

[35] M. Petraityte, A. Dehghantanha, and G. Epiphaniou, "Mobile phone forensics: an investigative framework based on user impulsivity and secure collaboration errors," in Contemporary Digital Forensic Investigations of Cloud and Mobile Applications. Elsevier, 2017, pp. 79–89.

[36] H. Dreger, A. Feldmann, M. Mai, V. Paxson, and R. Sommer, "Dynamic application-layer protocol analysis for network intrusion detection," in 15th USENIX security symposium. USENIX Association, 2006, pp. 257–272.

[37] G. Maier, R. Sommer, H. Dreger, A. Feldmann, V. Paxson, and F. Schneider, "Enriching network security analysis with time travel," in ACM SIGCOMM Computer Communication Review, vol. 38, no. 4. ACM, 2008, pp. 183–194.

[38] M. Saadeh, A. Sleit, M. Qatawneh, and W. Almobaideen, "Authentication techniques for the internet of things: A survey," in 2016 Cybersecurity and Cyberforensics Conference (CCC). IEEE, 2016, pp. 28–34.

[39] K. Shanmugasundaram, N. Memon, A. Savant, and H. Bronnimann, "Fornet: A distributed forensics network," in International Workshop on Mathematical Methods, Models, and Architectures for Computer Network Security. Springer, 2003, pp. 1–16.

[40] W. Wang and T. E. Daniels, "A graph based approach toward network forensics analysis," ACM Transactions on Information and System Security (TISSEC), vol. 12, no. 1, p. 4, 2008.

[41] T. V. Lillard, Digital forensics for network, Internet, and cloud computing: a forensic evidence guide for moving targets and data. Syngress Publishing, 2010.

[42] C. Prosise, K. Mandia, and M. Pepe, "Incident response & computer forensics," 2003.

[43] R. Geambasu, T. Bragin, J. Jung, and M. Balazinska, "On-demand view materialization and indexing for network forensic analysis." in NetDB, 2007.

[44] A. Singhal, C. Liu, and D. Wijesekara, "Poster: A logic based network forensics model for evidence analysis," in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM, 2015, pp. 1677–1677.

[45] M. Neugschwandtner, P. M. Comparetti, G. Jacob, and C. Kruegel, "Forecast: skimming off the malware cream," in Proceedings of the 27th Annual Computer Security Applications Conference. ACM, 2011, pp. 11–20.

[46] U. Bayer, P. M. Comparetti, C. Hlauschek, C. Kruegel, and E. Kirda, "Scalable, behavior-based malware clustering." in NDSS, vol. 9. Citeseer, 2009, pp. 8–11.

[47] T. Tafazzoli, E. Salahi, and H. Gharaee, "A proposed architecture for network forensic system in large-scale networks," arXiv preprint arXiv:1508.01890, 2015.

[48] L. Jiang, G. Tian, and S. Zhu, "Design and implementation of network forensic system based on intrusion detection analysis," in 2012 International Conference on Control Engineering and Communication Technology. IEEE, 2012, pp. 689–692.

[49] E. Oriwoh, D. Jazani, G. Epiphaniou, and P. Sant, "Internet of things forensics: Challenges and approaches," in 9th IEEE International Conference on Collaborative computing: networking, Applications and Worksharing. IEEE, 2013, pp. 608–615.

[50] M. H. Qasem and M. Qatawneh, "Parallel hill cipher encryption algorithm," International Journal of Computer Applications, vol. 179, no. 19, pp. 16–24, 2018.

[51] A. MacDermott, T. Baker, and Q. Shi, "Iot forensics: Challenges for the ioa era," in 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS). IEEE, 2018, pp. 1–5.

[52] V. R. Kebande and I. Ray, "A generic digital forensic investigation framework for internet of things (iot)," in 2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud). IEEE, 2016, pp. 356–362.

[53] L. Babun, A. K. Sikder, A. Acar, and A. S. Uluagac, "Iotdots: A digital forensics framework for smart environments," arXiv preprint arXiv:1809.00745, 2018.

[54] A. Nieto, R. Rios, and J. Lopez, "A methodology for privacy-aware iot-forensics," in 2017 IEEE Trustcom/BigDataSE/ICESS. IEEE, 2017, pp. 626–633.

[55] "Iot-forensics meets privacy: towards cooperative digital investigations," Sensors, vol. 18, no. 2, p. 492, 2018.

[56] T. Zia, P. Liu, and W. Han, "Application-specific digital forensics investigative model in internet of things (iot)," in Proceedings of the 12th International Conference on Availability, Reliability and Security. ACM, 2017, p. 55.

[57] K. Ruan, J. Carthy, T. Kechadi, and I. Baggili, "Cloud forensics definitions and critical criteria for cloud forensic capability: An overview of survey results," Digital Investigation, vol. 10, no. 1, pp. 34–43, 2013.

[58] C. Esposito, A. Castiglione, and K.-K. R. Choo, "Challenges in delivering software in the cloud as microservices," IEEE Cloud Computing, vol. 3, no. 5, pp. 10–14, 2016.

[59] A. Pichan, M. Lazarescu, and S. T. Soh, "Cloud forensics: Technical challenges, solutions and comparative analysis," Digital Investigation, vol. 13, pp. 38–57, 2015.

[60] K. Ruan, J. Carthy, T. Kechadi, and M. Crosbie, "Cloud forensics," in IFIP International Conference on Digital Forensics. Springer, 2011, pp. 35–46.

[61] M. Damshenas, A. Dehghantanha, R. Mahmoud, and S. bin Shamsuddin, "Forensics investigation challenges in cloud computing environments," in Proceedings Title: 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec). IEEE, 2012, pp. 190–194.

[62] S. Zawoad and R. Hasan, "Cloud forensics: a meta-study of challenges, approaches, and open problems," arXiv preprint arXiv:1302.6312, 2013.

[63] R. Hegarty, M. Merabti, Q. Shi, and B. Askwith, "Forensic analysis of distributed data in a service oriented computing platform," in proceedings of the 10th Annual Postgraduate Symposium on The Convergence of Telecommunications, Networking & Broadcasting, PG Net, 2009.

[64] M. Rajallah Asassfeh, M. Qatawneh, and F. Alazzeh, "Performance evaluation of blowfish algorithm on supercomputer iman1," International journal of Computer Networks and Communications, vol. 10, pp. 43–53, 03 2018.

[65] M. Alkhanafseh and M. Qatawneh, "A parallel chemical reaction optimization algorithm for max flow problem," International Journal of Computer Science and Information Security,, vol. 15, pp. 19–32, 06 2017.

[66] H. Harahsheh and M. Qatawneh, "Performance evaluation of twofish algorithm on iman1 supercomputer," International Journal of Computer Applications, vol. 179, pp. 1–7, 06 2018.

[67] A. Al-Shorman and M. Qatawneh, "Performance of parallel rsa on iman1 supercomputer," International Journal of Computer Applications, vol. 180, pp. 31–36, 04 2018.

[68] A. Eleyan and D. Eleyan, "Forensic process as a service (fpaas) for cloud computing," in 2015 European Intelligence and Security Informatics Conference. IEEE, 2015, pp. 157–160.

[69] J. Shah and L. G. Malik, "An approach towards digital forensic framework for cloud," in 2014 IEEE International Advance Computing Conference (IACC). IEEE, 2014, pp. 798–801.

[70] S. Zawoad, R. Hasan, and A. Skjellum, "Ocf: an open cloud forensics model for reliable digital forensics," in 2015 IEEE 8th International Conference on Cloud Computing. IEEE, 2015, pp. 437–444.

[71] M. Hogan, F. Liu, A. Sokol, and J. Tong, "Nist cloud computing standards roadmap," NIST special publication, vol. 35, pp. 6–11, 2011.

[72] R. Adams, "The advanced data acquisition model (adam): A process model for digital forensic practice," Ph.D. dissertation, Murdoch University, 2012.

[73] S. Zawoad, A. K. Dutta, and R. Hasan, "Seclaas: secure logging-as-a-service for cloud forensics," in Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security. ACM, 2013, pp. 219–230.

[74] T. Sang, "A log based approach to make digital forensics easier on cloud computing," in 2013 Third International Conference on Intelligent System Design and Engineering Applications. IEEE, 2013, pp. 91–94.

[75] P. M. Trenwith and H. S. Venter, "Digital forensic readiness in the cloud," in 2013 Information Security for South Africa. IEEE, 2013, pp. 1–5.

[76] R. Marty, "Cloud application logging for forensics," in proceedings of the 2011 ACM Symposium on Applied Computing. ACM, 2011, pp. 178–184.

[77] G. Grispos, T. Storer, and W. B. Glisson, "Calm before the storm: The challenges of cloud computing in digital forensics," International Journal of Digital Crime and Forensics (IJDCF), vol. 4, no. 2, pp. 28–48, 2012.

[78] "Computer Forensics: Network Forensics Analysis and Examination Steps,"https://resources.infosecinstitute.com/category/computerforensics/introduction/areas-of-study/digital-forensics/network-forensics-analysis-and-examination-steps/#gref., 2019.

[79] K. Ruan, "Cybercrime and cloud forensics: Applications for investigation," 2013.

[80] A. Haeberlen, "A case for the accountable cloud," ACM SIGOPS Operating Systems Review, vol. 44, no. 2, pp. 52–57, 2010.

[81] R. K. Ko, P. Jagadpramana, M. Mowbray, S. Pearson, M. Kirchberg, Q. Liang, and B. S. Lee, "Trustcloud: A framework for accountability and trust in cloud computing," in 2011 IEEE World Congress on Services. IEEE, 2011, pp. 584–588.

[82] S. Gokuldev and S. Leelavathi, "Hasbe: A hierarchical attribute- based solution for flexible and scalable access control by separate encryption/decryption in cloud computing," International Journal of Engineering Science and Innovative Technology (IJESIT), vol. 2, no. 3, 2013.

[83] S. Khan, A. Gani, A. W. A. Wahab, M. A. Bagiwa, M. Shiraz, S. U. Khan, R. Buyya, and A. Y. Zomaya, "Cloud log forensics: foundations, state of the art, and future directions," ACM Computing Surveys (CSUR), vol. 49, no. 1, p. 7, 2016.

[84] M. Aydin and J. Jacob, "A comparison of major issues for the development of forensics in cloud computing," in 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013). IEEE, 2013, pp. 77–82.